What are the determinants of DNA demethylation following treatment of AML cell lines and patient samples with decitabine?

by

Robert John Hollows

This project is submitted in partial fulfilment of the requirements for the award of the MRes in Biomedical Research

School of Cancer Sciences

College of Medical and Dental Sciences

University of Birmingham

September 2012

Abstract

Aberrant DNA methylation of CpG sites has been linked to the aetiology and pathogenesis of various malignancies including acute myeloid leukemia (AML). Decitabine is a drug which has been shown to reduce methylation levels, and is being increasingly explored as an agent for treating cancer. However, the determinants of demethylation caused by decitabine are not completely understood.

The purposes of this study were to investigate the determinants of the demethylation observed following treatment of AML samples with decitabine, and to explore whether these determinants could explain the variation in demethylation across two AML cell lines and eight primary cultures from AML patients.

The results showed considerable variation in the level of demethylation between samples. Within samples, CpG demethylation was found to vary according to CpG location, CpG density, proximity to a CpG island and pre-treatment methylation levels. Multivariate regression analysis showed that the principal determinant of demethylation at an individual CpG site was the pre-treatment methylation level. However, the analysis also showed that the determinants identified were in themselves insufficient to explain all of the variation in demethylation observed across study samples.

Acknowledgements

I would like to thank Professor Ciaran Woodman, Dr Sarah Leonard and Dr Wenbin Wei for their valuable assistance and guidance with this project.

Table of contents

List of figures

List of tables

1. Introduction

This project examined the effect of a global demethylating agent, decitabine (DAC), on the methylation profiles of acute myeloid leukemia cells taken from eight paediatric patients and two cell lines.

1.1 Acute myeloid leukemia

Acute myeloid leukemia (AML) is a cancer of blood cells which typically, but not exclusively, affects older people and is caused by the proliferation of abnormal white blood cells which interfere with the production and function of normal blood cells.

AML has a heterogeneous genetic basis, and certain known chromosomal abnormalities are used to classify patients into favourable, intermediate and unfavourable prognostic groups[1]. Furthermore there are a small number of single gene mutations which have been discovered to be markers of good (NPM1) and bad (FLT3) prognosis[2]. It is also believed that epigenetic modifications, in particular abnormal levels of methylation, to the underlying DNA have a causal implication in AML initiation[3].

1.2 DNA methylation

Methylation is a chemical process which modifies DNA in such a way that the change can be inherited (via cell division) without changing the underlying DNA sequence. It typically affects the cytosine element of a cytosine/guanine di-nucleotide (referred to as a "CpG site"), whereby the cytosine is chemically tagged by a methyl group.

Methylation has a key role in the normal regulation of gene activity - typically methylation of a gene's promoter will result in repression of that gene, whereas methylation of CpG sites within the gene body itself can be associated with gene activation[4].

Maintenance of methylation levels is controlled by a family of enzymes known as DNA methyltransferases (DNMTs). Once a piece of DNA has been methylated, it can be bound by methyl-binding proteins, which in turn recruit enzymes that modify the structure of the surrounding chromatin in such a way that results in repressed gene expression[5].

1.3 Cancer and methylation

It is unsurprising that abnormal levels of DNA methylation have been linked to a number of cancers, including AML[6]. This typically takes the form of abnormally high (hyper-) methylation of the promoters of tumour suppressor genes, thereby causing the affected genes to become inactive, and thus facilitating loss of control over cell proliferation, potentially leading to cancer. CDKN2B and CEBPA are two examples of tumour suppressor genes which are silenced in AML as a result of abnormal methylation[7].

1.4 Decitabine (DAC)

As mentioned above, methylation levels across the genome are controlled by DNMTs. Hence, inhibiting the activity of DNMTs in cancerous cells is a potential means of reversing abnormally high methylation and thereby restoring the functionality of affected tumour suppressor genes. This is how DAC works via its effect on one of the DNMTs known as DNMT1[8].

The precise way in which DAC achieves this is not fully understood, but in broad terms it manages to incorporate itself into DNA and then bond with and trap DNMT1, thus inhibiting its activity[9]. This leads to what is known as "passive" demethylation whereby the high level of hyper-methylation becomes diluted following cell replication and division. DAC has been approved for the treatment of myelodysplastic syndrome, which is a preleukemic bone marrow disorder.

1.5 Genome-wide DNA methylation profiling

Over recent years there have been considerable technological developments enabling more efficient and accurate DNA methylation profiling at individual base-pair resolution on a genome-wide scale[10]. In particular, microarrays which use DNA hybridisation techniques have been adapted for this purpose. Furthermore, next generation sequencing technologies are also now beginning to be adopted.

Before the profiling by microarray or other technology takes place, a special preparatory step is generally required to be applied to the DNA because methylated and unmethylated cytosines are indistinguishable by these technologies[10]. The three key techniques used for this are:

1. digestion with restriction enzymes which are sensitive to methylation status;
2. affinity enrichment using antibodies which are specific for methylated cytosines; and
3. bisulphite conversion whereby DNA is treated with sodium bisulphite which causes unmethylated cytosines to be converted to uracil whilst methylated cytosines are unaltered.

This project was based on the results of DNA methylation profiling performed using the Infinium HumanMethylation450 BeadChip produced by Illumina Inc. This is an array-based technology which analyses DNA samples which have been bisulphite converted.

1.6 Aims of project

Whilst DAC is known to be a demethylating agent, the extent to which it demethylates individual CpG sites is not uniform across the whole human genome. Hence, the key aims of this project were to investigate, on a genome-wide scale, the determinants of demethylation

observed following treatment of AML samples with DAC, and to explore whether these determinants could explain the variation in demethylation across samples. For this purpose, DNA samples had previously been taken from eight paediatric patients suffering from acute myeloid leukemia and also two AML cell lines. The methylation profiles of these samples had been measured both before and after treatment with DAC using the Illumina array, and these values were compared across individual CpG sites.

The most directly comparable item of previous research in this area[11] indicated that the level of demethylation of individual CpG sites following the administration of DAC to AML cell lines appears to be highly dependent on the initial (pre-treatment) level of methylation - i.e. demethylation is greater for CpG sites with a higher pre-treatment level of demethylation, and vice-versa. The same research also showed that CpG sites in so-called CpG islands[12] experienced lower demethylation than those outside such islands. However, this research was based on an older version of the Infinium technology which has since been upgraded considerably. This project uses the new technology (details in section 2), which allows much greater refinement of the analysis. Furthermore this project also uses patient samples as well as cell lines, and hence is based on data which has been extracted from sources which are closer to actual clinical reality.

2. Data and methods

Details of the data and methods used for this project are set out in the sections below. The experimental work described in sections 2.1 to 2.3 was performed by other people. The analysis of the resultant data carried out for this project is described in sections 2.4 and 2.5.

2.1 Sample data

The analysis was based on samples taken from eight, untreated paediatric patients with AML and from two AML cell lines (HL60 and KGIA). These samples were processed for methylation analyses between Autumn and October 2011 within the School of Cancer Sciences at the University of Birmingham. Patient samples are referred to as APAL, CBUN, DONCO, GALWIL, NALE, RILNIA, ROSBOS and ZPEA. Primary cultures were established from AML mononuclear cells isolated from bone marrow cells of the eight patients. There were no technical replicates for any of the samples.

In order to assess the impact of treatment on methylation status, each sample was treated in vitro with 0.05µM of DAC. Genome-wide methylation analysis was performed on each of these samples both before treatment and five days after treatment. The combination of treatment with 0.05µM of DAC and subsequent measurement after 5 days was found to result in the greatest decrease in methylation levels by using pyrosequencing on four sample genes.

This analysis forms part of a larger project looking into the impact of treatment with DAC on AML cell viability and linking the results to associated changes in gene expression. Initial results of viability assays indicate that samples from seven of the eight patients experienced reduced viability following treatment with DAC, with APAL being the one exception.

2.2 Methylation profiling

The genome-wide methylation analysis was performed using the Infinium HumanMethylation450 BeadChip produced by Illumina Inc. This is described in detail in the paper by Bibikova et al[13]. Briefly, the Infinium array allows high resolution interrogation of over 480,000 individual CpG sites for up to 12 samples simultaneously. It is based on a combination of Infinium I and Infinium II techniques, which both analyse bisulphite-converted DNA.

The Infinium I technology works by having two "beads" for each targetted CpG site which are used to hybridise with the sample DNA. For any particular CpG site, one of the beads is designed to hybridise with an unmethylated cytosine, whilst the other will hybridise with a methylated cytosine. The methylation proportion for the CpG site is the ratio of the number of hybridisations with the latter to the total number of hybridisations, with each hybridisation being recorded by detection of a (green) fluorescently labelled nucleotide.

The Infinium II technology uses only one bead for each CpG site. The methylation status is determined by single-base extension, whereby a red fluorescently labelled adenine will hybridise to an unmethylated (thymine) locus and a green fluorescently labelled guanine will hybridise to a methylated (cytosine) locus. The methylation proportion for a CpG site is then determined as the ratio of the amount of green fluorescence detected divided by the total of red and green fluorescence detected.

Figure 1 below is a schematic of the analysis process taken from the Bibikova paper.

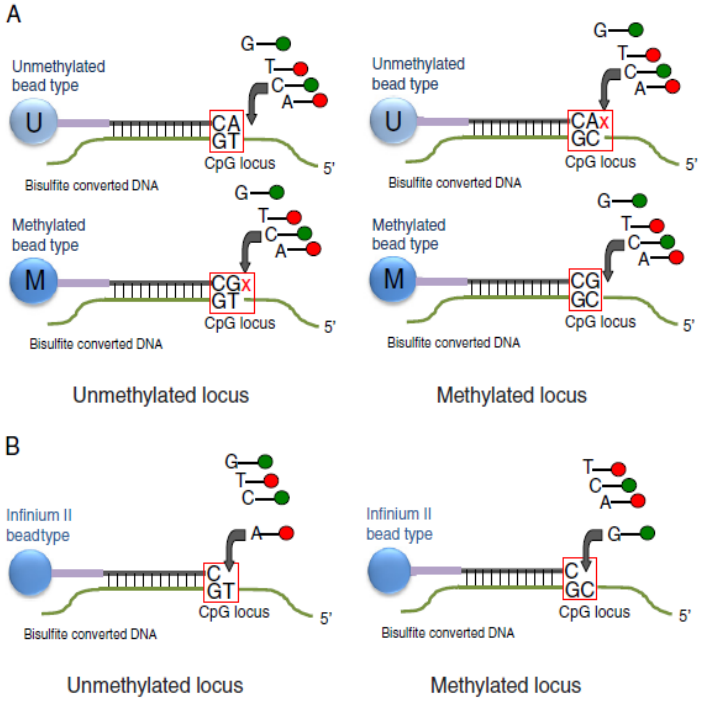


Figure 1 - schematic of (A) Infinium I and (B) Infinium II technology

To facilitate downstream analysis of the results, each CpG site is associated with two separate labels. The first relates to its location relative to the nearest gene (hereafter referred to as "gene location"), and the second its location relative to the nearest CpG island ("CpG island region"). Tables 1 and 2 show descriptions of the classifications used in the different labels.

Gene location	Description
Body	Region between 3'UTR and 1st exon
1st exon	1st exon
3'UTR	Untranslated region at the 3' end
5'UTR	Untranslated region at the 5' end
TSS1500	Region between 200 and 1,500 bases upstream of the transcriptional start site
TSS200	Region from the transcriptional start site to 200 bases upstream of this
Sites which do not fall into any of the above categories are unlabelled. For the purposes of this project, these have been called "intergenic".	

Table 1 - list of gene location labels used to categorise individual CpG sites

CpG island region	Description
Island[12]	>=500bp region with more than 50% GC composition and CpG observed/expected ratio of at least 60%
North shore[14]	2kb region upstream of an island
South shore	2kb region downstream of an island
North shelf	2kb region upstream of a north shore
South shelf	2kb region downstream of a south shore
Sites which do not fall into any of the above categories are unlabelled. For the purposes of this project, these have been called "ocean".	

Table 2 - list of CpG island regions used to categorise individual CpG sites

A schematic representation of these labels, taken from the Bibikova paper, is shown below in figure 2:

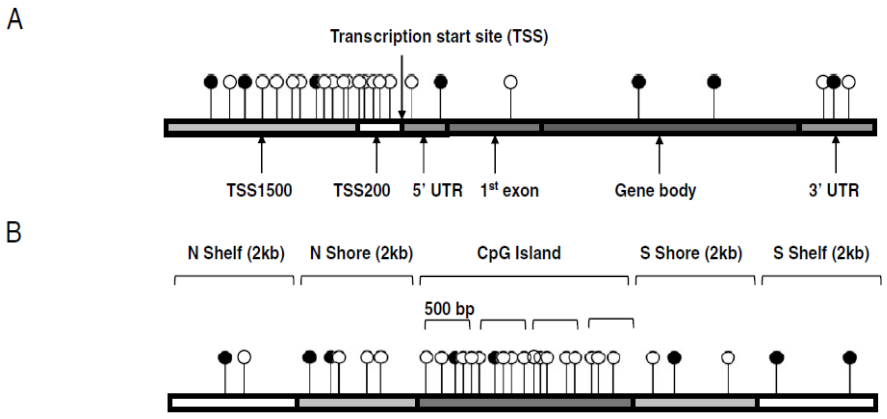


Figure 2 - schematic of (A) gene locations and (B) CpG island regions

2.3 Raw data processing

In order to calculate the methylation status for each CpG site for each sample (both before and after treatment with DAC), the raw data produced by the Infinium array were processed using the Genome Studio (version 1.8) software produced by Illumina. No normalisation or background control adjustment was applied.

In order to check the accuracy of the output, pyrosequencing was also performed, for all eight patient samples, on four selected genes. The pyrosequencing analysis confirmed the Illumina array results for all samples and candidate genes within a reasonable margin of error, which was estimated to be of the order of 5%.

The key output provided by the Illumina array is the so-called beta-value for each CpG site. For any particular CpG site, this represents the average level of methylation detected across all the probes for that site, and ranges from 0% (totally unmethylated) to 100% (totally methylated). The beta value data were the starting point for this project.

2.4 Initial filtering of raw beta values

Whenever the Infinium array is used, it is possible that it may not function properly for some of the CpG sites and, therefore, may not adequately measure their methylation statuses. The extent of any malfunctioning can be investigated for each CpG site by looking at the so-called "detection p-values" output by the Genome Studio software. The detection-p-value for each CpG site is calculated by comparing each of the measured methylated probe and demethylated probe intensities with the intensity distribution of negative control probes. For any given CpG site, a high p-value indicates that there is a potential problem with the probes for that site.

In order to filter out those sites for which the beta values calculated by Genome Studio were deemed to be too unreliable for the purposes of this study, software called IMA (Illumina Methylation Analyzer)[15], developed for the programming language R, was used. Any sites for which at least 75% of samples had a detection p-value greater than 5% were filtered out using this software. All other sites which passed this test were used in the next stage of analysis.

2.5 Analysis of filtered beta values

The key part of the analysis was to investigate the factors which appear to impact on the change in methylation level caused by treatment with DAC.

The change in methylation level for an individual CpG site can be calculated in two different ways. Firstly, the arithmetical difference between average beta values before and after treatment could be used - for example, if the pre-treatment average beta value is 80% and the post-treatment value is 60%, then this would represent a reduction of 20%. This type of change is hereafter referred to as the "absolute" change. An alternative way of calculating the change is to look at the proportionate change - using the same example, the proportionate change (reduction in this case) would be 25% (because 20% is 25% of 80%).

Previous literature (e.g. the Hagemann paper[11]) has tended to analyse absolute rather than proportionate changes. A key disadvantage of using proportionate changes is that, when applied to small initial values, large proportionate changes still result in small absolute changes which are potentially within the margin of error of the measurement accuracy of the Infinium array. Hence analysis based on proportionate changes could be subject to distortion caused by measurement inaccuracy.

Based on an analysis of completely separate data which contained two sets of two technical replicates, this study has found that 99.5% of all pairs of replicate beta value measurements using the Illumina array had an absolute difference of no more than 10%. However, if proportionate differences are considered, then the maximum difference covering 99.5% of all pairs is much higher at 67%. At the lower confidence level of 95%, the upper bounds are 5% (absolute differences) and 30% (proportionate).

Given the potential for introducing distortions if proportionate changes are used, this study has concentrated mainly on absolute changes, and these are what are used in the text unless otherwise stated. However, proportionate changes do also need to be considered, as they may result in different conclusions being reached. Hence, results using proportionate changes are also included in various parts of section 3.

Four key factors were identifed as being potentially associated with the changes in beta value. These are:

- starting (pre-treatment) beta value
- gene location
- CpG island region
- CpG density

The first three of these variables are all directly available from the Genome Studio output. The fourth was calculated using a piece of code written in R (using the BSgenome package for genome information). For this purpose, CpG density was calculated for each CpG site across a 500bp region centred on the site. The total number of CG dinucleotides in the 500bp region was used as the measure of CpG density.

The associations of the four factors identified above with the change in beta value were then analysed using a mixture of programs written in R and Microsoft Excel. In particular, both linear and logistic regression analyses were performed in R (using the lm and glm functions respectively), and kernel density plots were created in R using the density function. For the logistic regression analysis, Nagelkerke's R^2 measure of association[16] was calculated using the fmsb package.

3. Results

The results of the analysis are set out in the paragraphs below.

Orientation of analysis

As previously stated, the aim of this study is to attempt to explain the variation in the level of demethylation of CpG sites in AML cell lines and patient samples following treatment with decitabine. The results are presented in the following order:

Firstly, section 3.1 analyses the key features of the Illumina methylation array which are intrinsic to the human genome and hence are the same for all samples. In particular the analysis investigates the distribution of CpG sites interrogated by the array, and how this distribution varies by the three factors, gene location, CpG island region and CpG density (as described in section 2).

Section 3.2 then analyses how the distribution of CpG sites varies according to pre-treatment methylation levels. For this purpose, CpG sites are grouped into 10% methylation bands (i.e. 0% to 10%, 10% to 20%, etc). At this point differences between the cell lines and patient samples will emerge as their pre-treatment methylation profiles are not identical. The way in which pre-treatment methylation levels vary across the other three intrinsic factors is also analysed.

Sections 3.3, 3.4 and 3.5 analyse the levels of demethylation in all the cell lines and patient samples following treatment with DAC. Section 3.4 investigates how the extent of demethylation varies across the three intrinsic factors, and then section 3.5 extends this analysis by comparing demethylation levels across pre-treatment methylation groups.

Finally, in section 3.6 the results of regression analyses are set out, which attempt to explain how the level of demethylation across CpG sites varies, using the three intrinsic factors and pre-treatment methylation levels as explanatory variables.

3.1 Variation in distribution of CpG sites

The Illumina array provides methylation data for 485,577 CpG sites. After the initial filtering of the most unreliable probes (as described in Section 2), 479,629 CpG sites remained for further analysis. The following paragraphs show how the distribution of these sites varies across the different gene locations, CpG island regions and according to CpG density (as measured by the number of CpG dinucleotides in the 500bp window centred on each site, described in section 2).

3.1.1 Distribution of CpG sites across gene locations

Figure 3 below illustrates how the proportions of CpG sites on the array vary across the different gene locations.

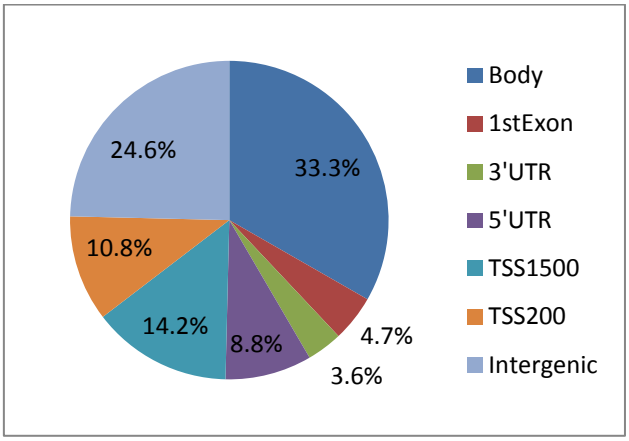


Figure 3 - distribution of CpG sites across gene locations

Gene bodies contain the highest number of CpG sites, with 33.3% of the total number. Intergenic sites are the next highest with 24.6%. 3'UTRs have the lowest number of CpG sites (3.6%), followed by 1st exons (4.7%).

3.1.2 Distribution of CpG sites across CpG island regions

Figure 4 below shows how the proportions of CpG sites on the array vary across the different CpG island regions.

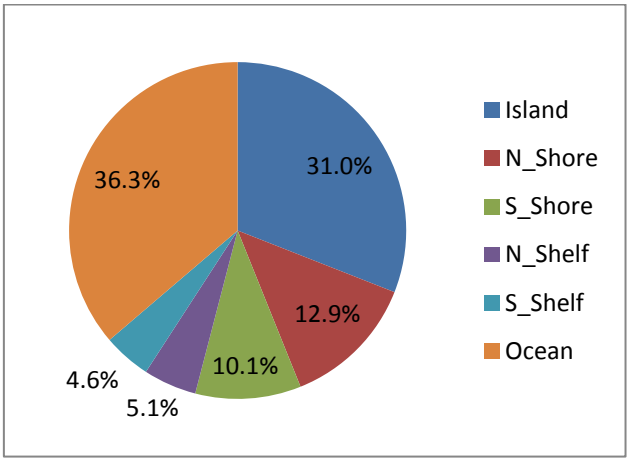


Figure 4 - distribution of CpG sites across CpG island regions

Ocean regions have the largest number of CpG sites, with 36.3% of the total, followed by CpG islands with 31.0%. North and south shelves have the lowest numbers of CpG sites with 5.1% and 4.6% respectively.

3.1.3 Distribution of CpG sites across areas of different CpG density

As described in section 2, some regions of the human genome are sparsely populated with CpG sites, whereas others are densely populated. Furthermore, CpG density has been shown to be a predictor or methylation changes when primary cells have been infected with oncogenic viruses[17]. Therefore, this section analyses how the numbers of CpG sites vary across regions of different CpG density.

For each CpG site, the CpG density of the region in which the site is located has been measured as the number of CpG dinucleotides which are found in the 500bp window centred

on the site. This measure varies between 1 and 109. Figure 5 below shows how the numbers of CpG sites vary across areas of different CpG density.

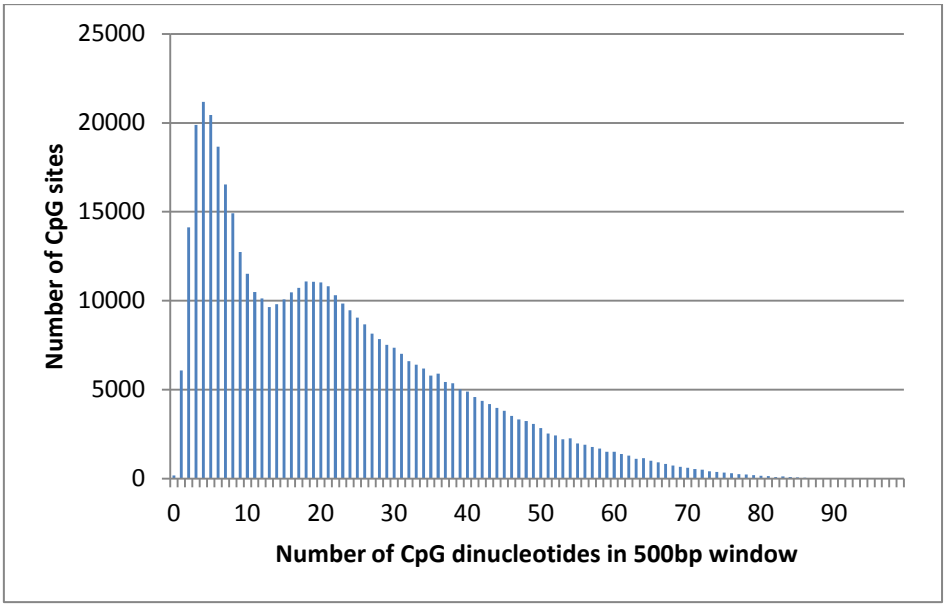


Figure 5 - distribution of CpG sites across areas of different CpG density

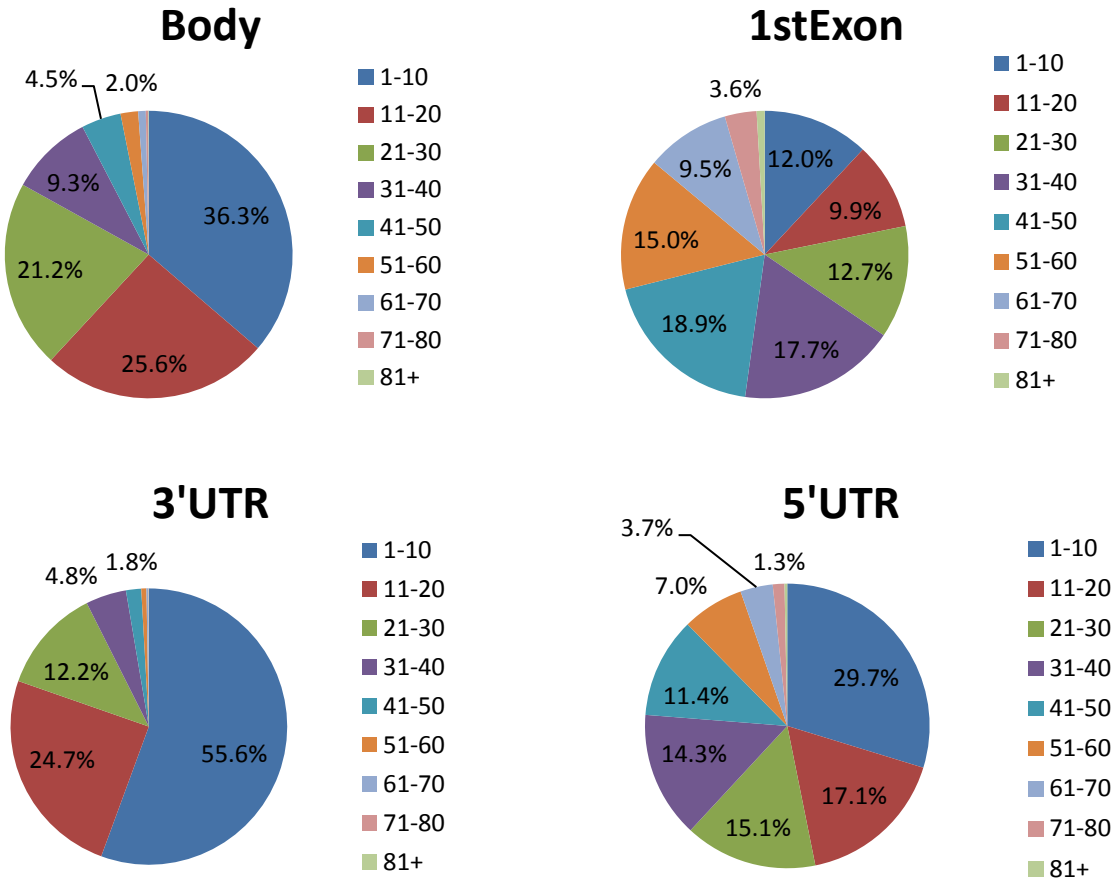
Most CpG sites on the array are located in regions of low CpG density. There is only a small proportion of CpG sites in areas of high CpG density.

The next two sections consider how the numbers of CpG sites in each gene location and island region vary across areas of different CpG density. For this purpose CpG counts are grouped into tens (1 to 10, 11 to 20, etc), except for the final group which covers all values above 80.

3.1.4 Distribution of CpG sites for each gene location, stratified by CpG density

Figure 6 below shows, for each gene location type, how the proportions of CpG sites vary across the different CpG density groups.

It is very rare for CpG sites in any gene location to be in an area of very high CpG density (CpG count above 70). Gene bodies, 3'UTRs, 5'UTRs, TSS1500s and intergenic regions all have a large proportion of sites in areas of low CpG density (CpG count of 30 or less) - for example, 93% of CpG sites located in 3'UTRs are found in such areas . 1st exons and TSS200s have lower proportions of sites in areas of low CpG density.



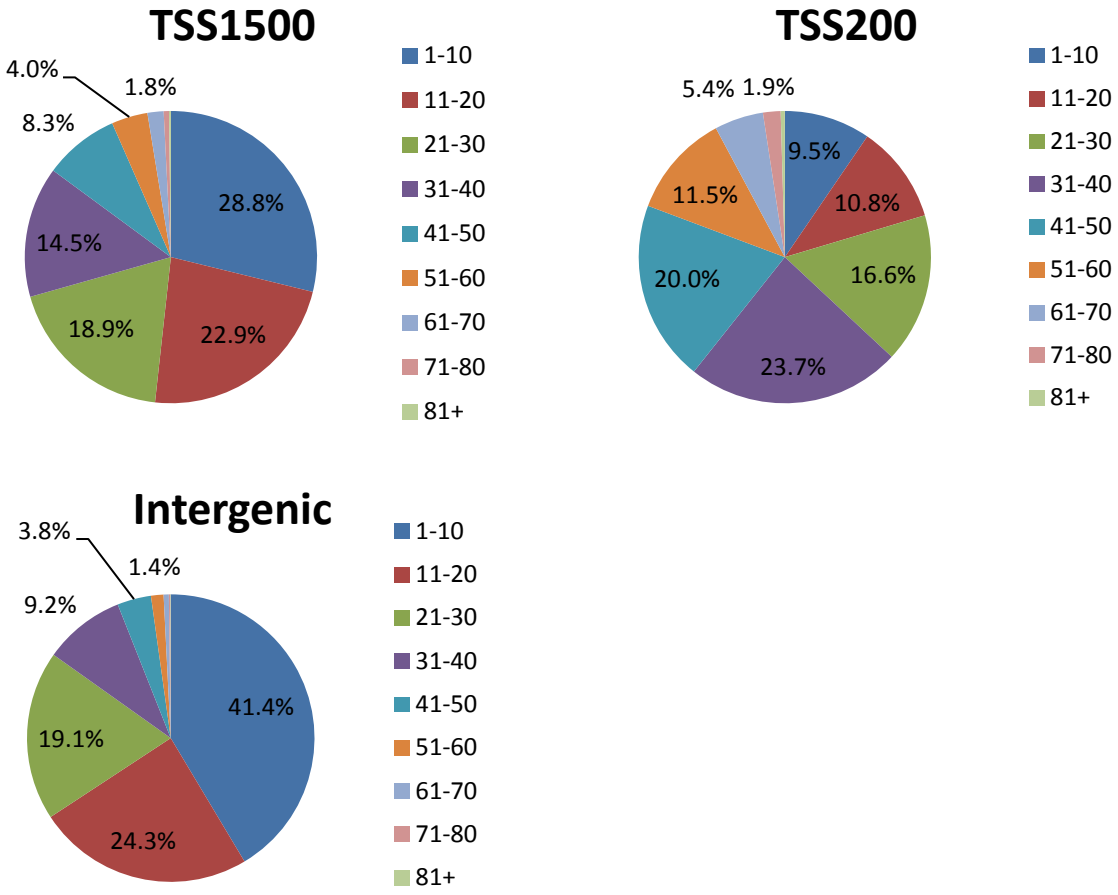


Figure 6 - variation in distributions of CpG sites for each gene location across different CpG density groups

3.1.5 Distribution of CpG sites for each CpG island region, stratified by CpG density

Figure 7 below shows, for each CpG island region type, how the proportions of CpG sites vary across the different CpG density groups.

For all CpG island regions, the proportion of sites in very high density areas (CpG count greater than 70) is very small. CpG sites in ocean regions, north shelves and south shelves are mainly located in areas of low CpG density - for example, 99% of CpG sites in north shelves are located in areas with a CpG count of 30 or less. North and south shores have higher proportions of sites in higher density areas, and islands have much higher proportions - 77% are located in areas with a CpG count of more than 30.

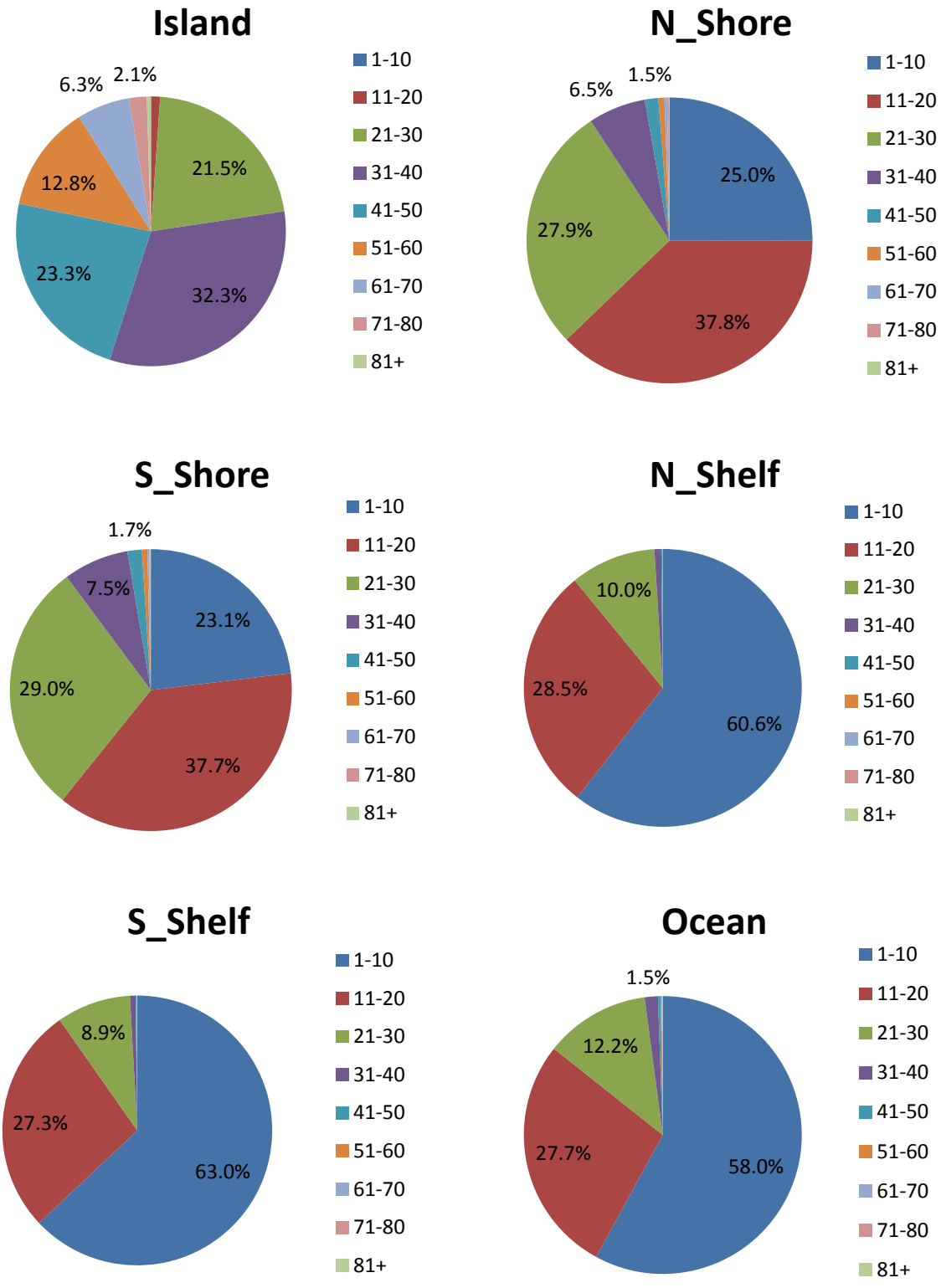


Figure 7 - variation in distributions of CpG sites for each CpG island region across different CpG density groups

3.1.6 Summary of key points

The analysis above shows that the distribution of CpG sites varies across gene locations and CpG island regions. Most sites are located in areas of low CpG density, with only a very small proportion found in areas of high density. Within gene locations and CpG island regions there is variation in the proportion of sites in areas of different CpG density.

3.2 Variation in pre-treatment methylation levels across CpG sites

The three factors considered in section 3.1 are, of course, consistent across cell lines and patient samples. However, this is not the case for the methylation profile observed before treatment with DAC

The kernel density plot in figure 8 below shows, for each cell line and patient sample, how pre-treatment methylation levels vary across all CpG sites on the array (note - for density plots the Y-axis units are arbitrary such that the area under each curve equals unity).

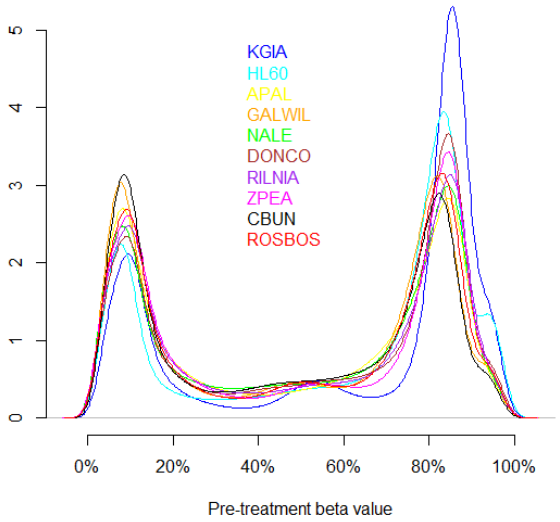


Figure 8 - pre-treatment methylation profiles for each cell line and patient sample

All the cell lines and patient samples have a bi-modal distribution of methylation levels, with one peak at around 10% methylation and another, higher peak at around 85%. However, there are some differences between the profiles. These differences are summarised in table 3 below, which shows the average pre-treatment beta values for each cell line and patient sample.

HL60	KGIA	APAL	GALWIL	NALE	DONCO	RILNIA	ZPEA	CBUN	ROSBOS
59.7%	61.8%	50.6%	50.2%	52.0%	54.1%	51.7%	51.2%	48.6%	51.2%

Table 3 - average pre-treatment beta values for each cell line and patient sample

In particular, the average values for the two cell lines are more than 10% higher than those for any of the eight patient samples. Across the latter, the range is from 48.6% (CBUN) to 54.1% (DONCO).

3.2.1 Average pre-treatment methylation levels stratified by gene location

Figure 9 below shows, for each cell line and patient sample, how the average pre-treatment beta values vary across different gene locations.

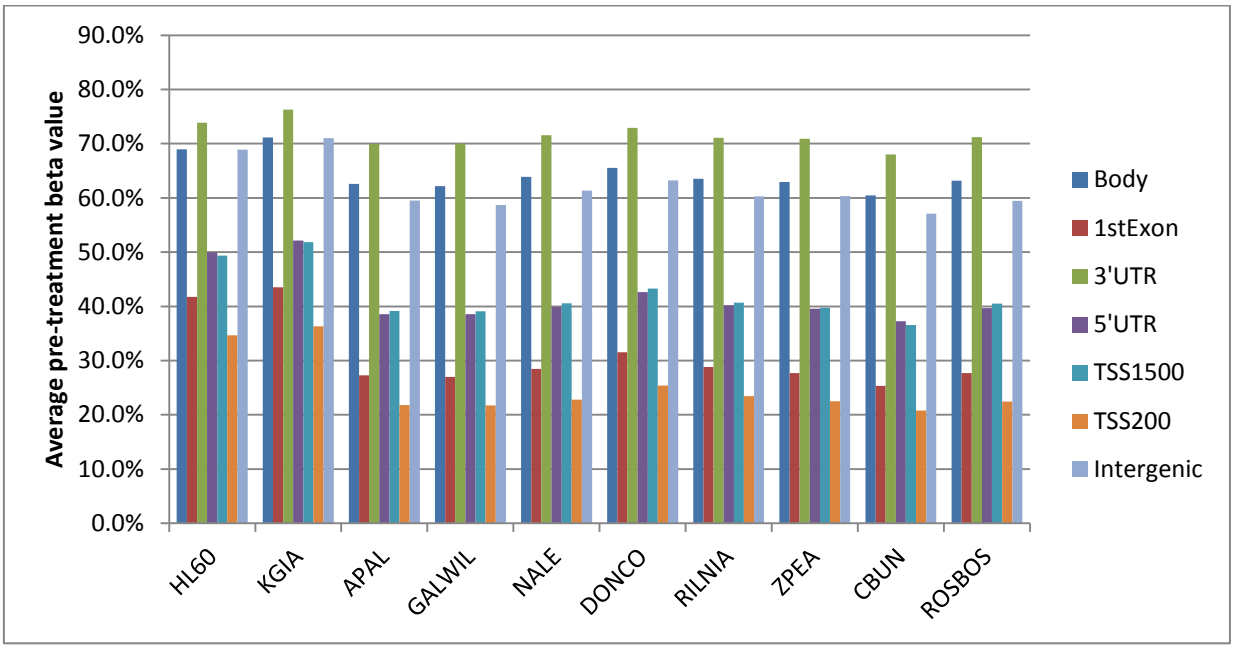


Figure 9 - average pre-treatment beta values for each gene location

Across all the cell lines and patient samples, 3'UTRs have the highest average pre-treatment beta values, followed by bodies and intergenic regions. Conversely, TSS200s have the lowest values, followed by 1st exons. Across all gene locations, the two cell lines have higher values than all of the patient samples.

3.2.2 Average pre-treatment methylation levels stratified by CpG island region

Figure 10 below shows, for each cell line and patient sample, how the average pre-treatment beta values vary across different CpG island regions.

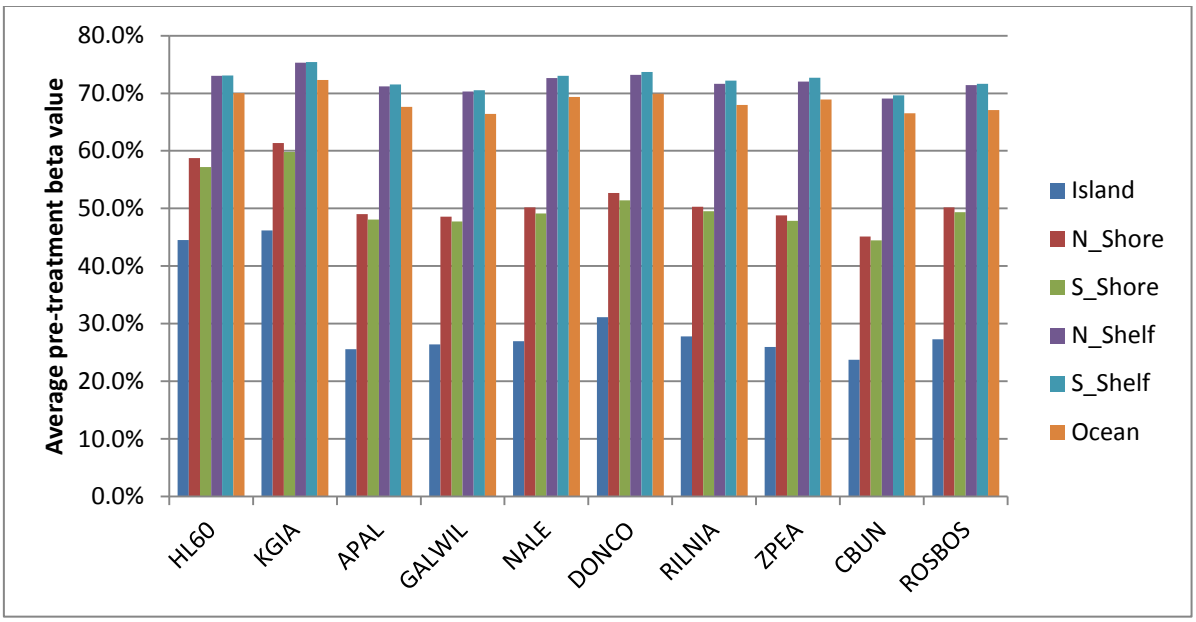


Figure 10 - average pre-treatment beta values for each CpG island region

For all cell lines and patient samples, north and south shelves have the highest average pre-treatment beta values, followed closely by ocean regions. Conversely, islands have the lowest values. Again the two cell lines have the highest values across all regions, with this difference being most pronounced for CpG islands.

3.2.3 Average pre-treatment methylation levels stratified by CpG density

Figure 11 below shows, for each cell line and patient sample, the average pre-treatment beta values stratified by CpG density group.

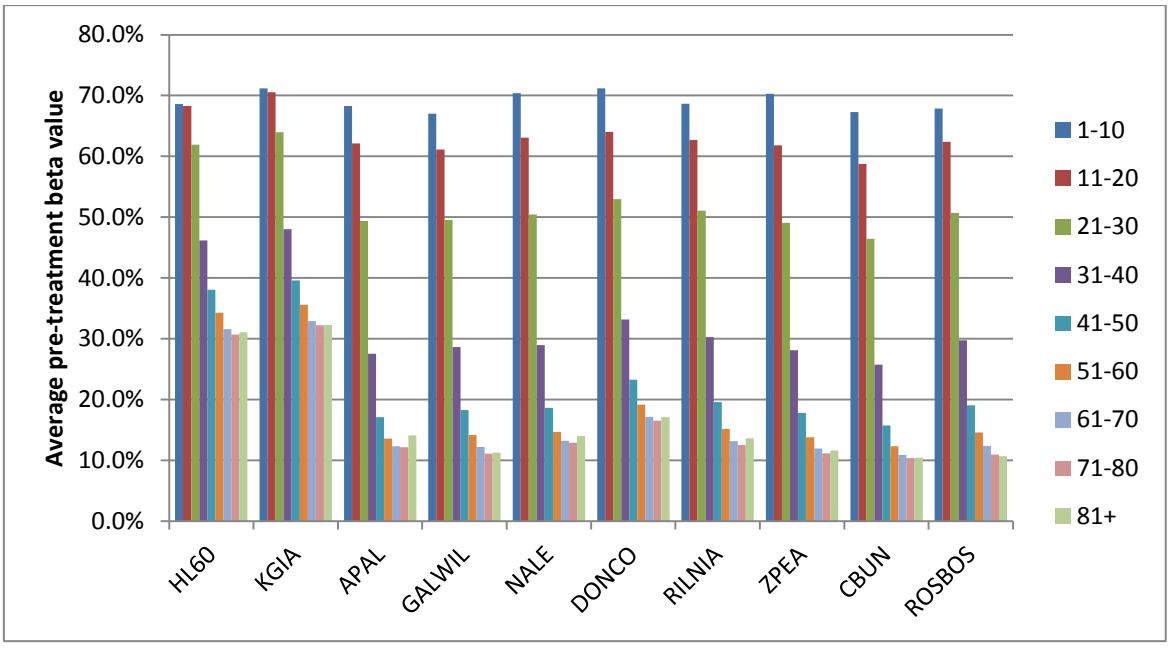


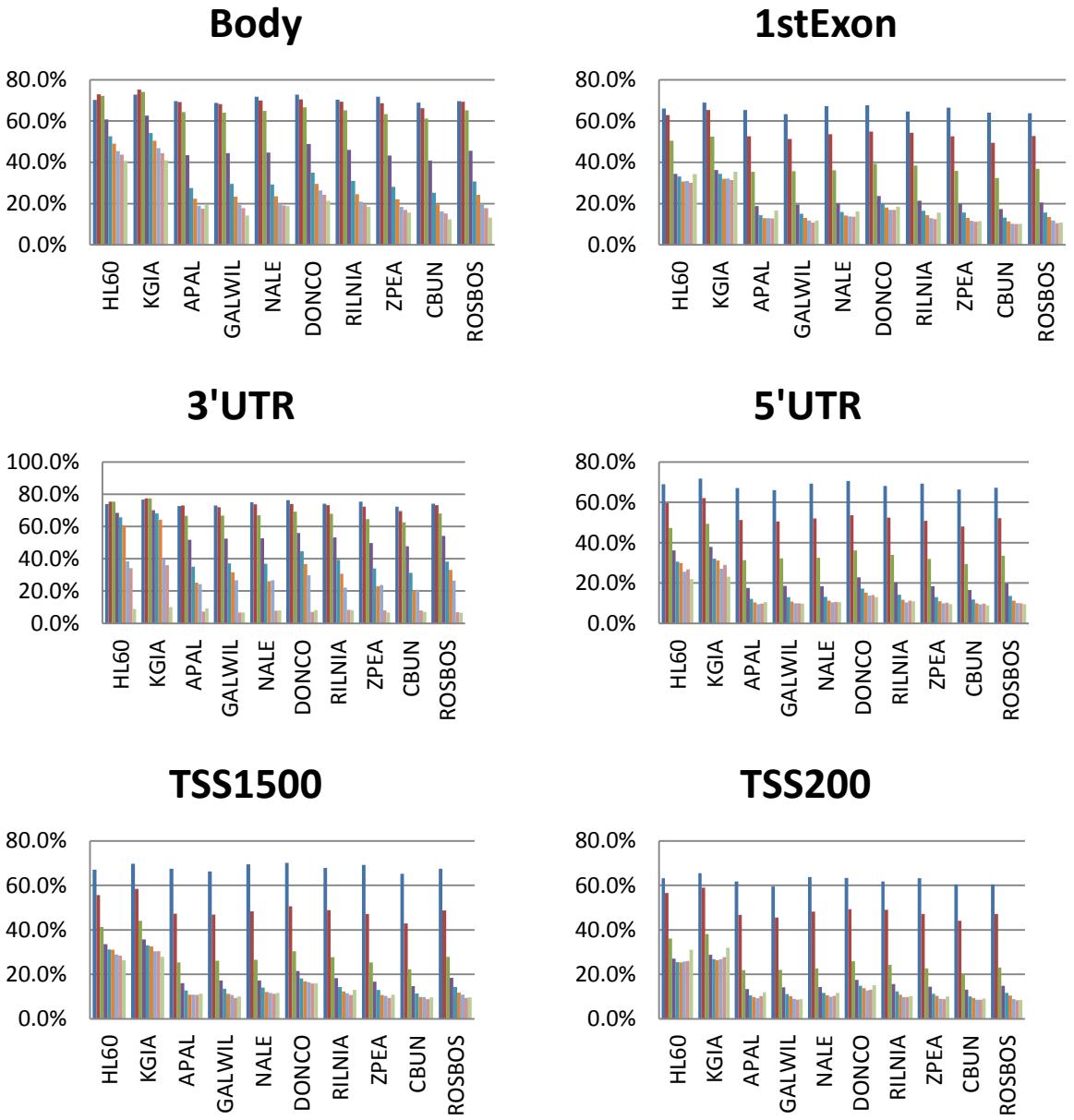
Figure 11 - average pre-treatment beta values for each CpG density group

Across all cell lines and patient samples, there is a clear trend for reducing pre-treatment methylation levels as CpG density increases. The pattern is, however, less pronounced in the two cell lines, for which average beta value remains above 30% for all CpG density groups.

3.2.4 Average pre-treatment methylation levels for each gene location stratified by CpG density

The next two sections consider how the average pre-treatment methylation levels of CpG sites in each gene location and island region vary across areas of different CpG density.

Firstly, figure 12 below shows, for each cell line and patient sample, the average pre-treatment beta values (y-axis) for each gene location, stratified by CpG density group. The trend for reducing average pre-treatment methylation level as CpG density increases can be seen for all gene locations. However it is less pronounced for bodies and intergenic regions, and more pronounced for 3'UTRs.



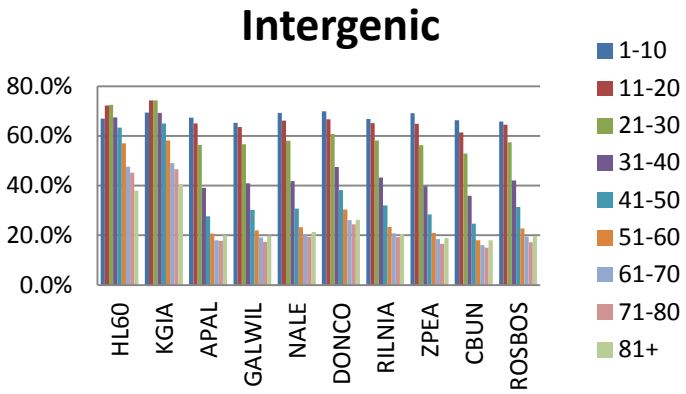
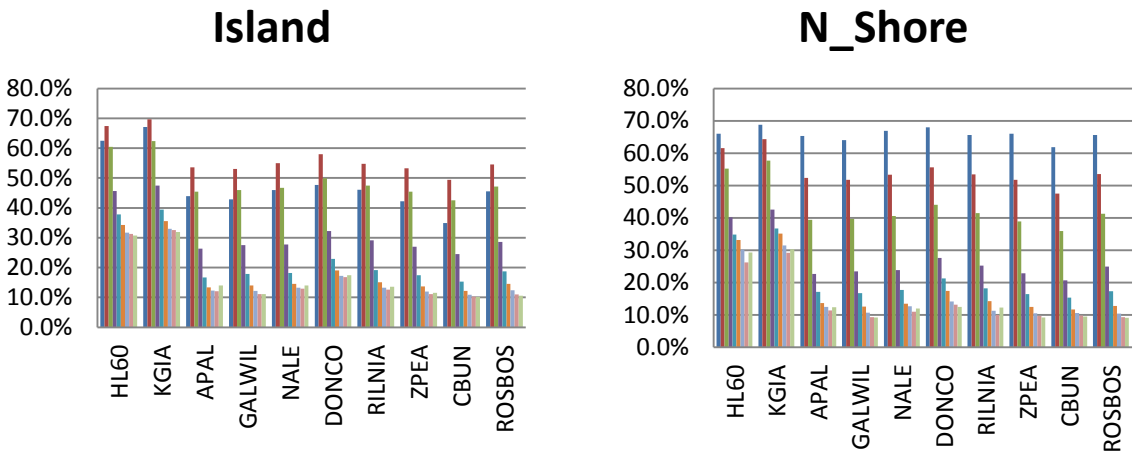


Figure 12 - average pre-treatment beta values for each gene location, stratified by CpG density groups

3.2.5 Average pre-treatment methylation levels for each CpG island region stratified by CpG density

Figure 13 below shows, for each cell line and patient sample, the average pre-treatment beta values (y-axis) for each CpG island region, stratified by CpG density group. The trend for reducing average pre-treatment methylation level as CpG density increases can again be seen for all CpG regions, although this trend is not consistent, in particular at the very highest CpG densities.



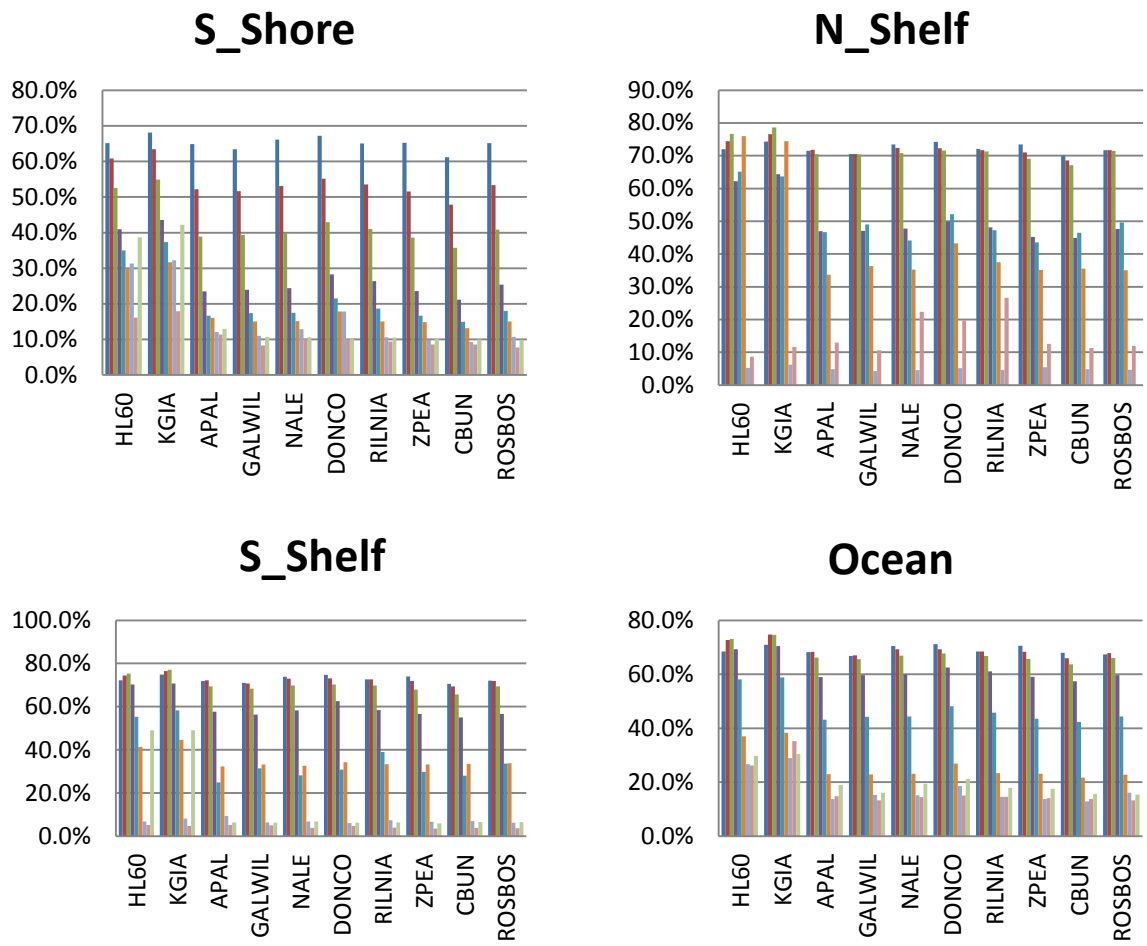


Figure 13 - average pre-treatment beta values for each CpG island region, stratified by CpG density groups (legend as Figure 12)

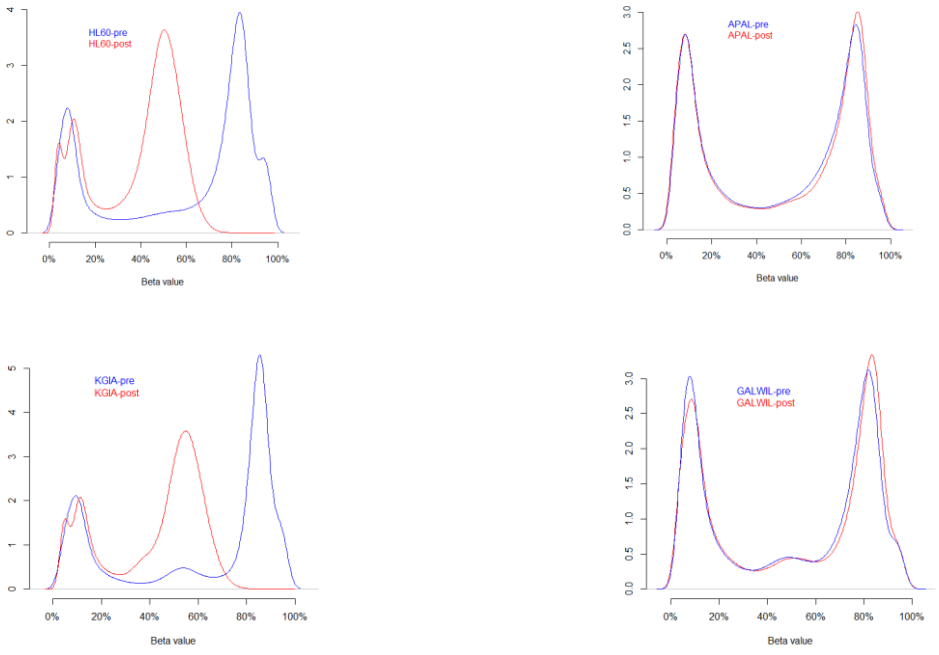
3.2.6 Summary of key points

The analysis of pre-treatment methylation levels shows that both cell lines and all patient samples have a bi-modal distribution of pre-treatment beta values, with one peak at around 10% methylation and another higher peak at around 85%. However the distributions vary between cell lines and patient samples. There is also variation across gene locations and CpG island regions and with CpG density. Sites in areas of low CpG density tend to be highly methylated, whereas those in areas of high CpG density tend to have low methylation levels.

3.3 Changes in methylation levels following treatment with DAC

The previous sections have shown how the distribution of CpG sites varies across gene locations and CpG island regions and with CpG density, and also how pre-treatment methylation levels vary across samples and according to these three intrinsic factors. This section now looks at the changes in methylation profile after treatment with DAC.

The extent of demethylation following treatment with DAC varied across the cell lines and patient samples. This variation is illustrated in figure 14 below, which shows, for the different cell lines and patient samples, the distributions across all CpG sites of beta values both before and after treatment with DAC.



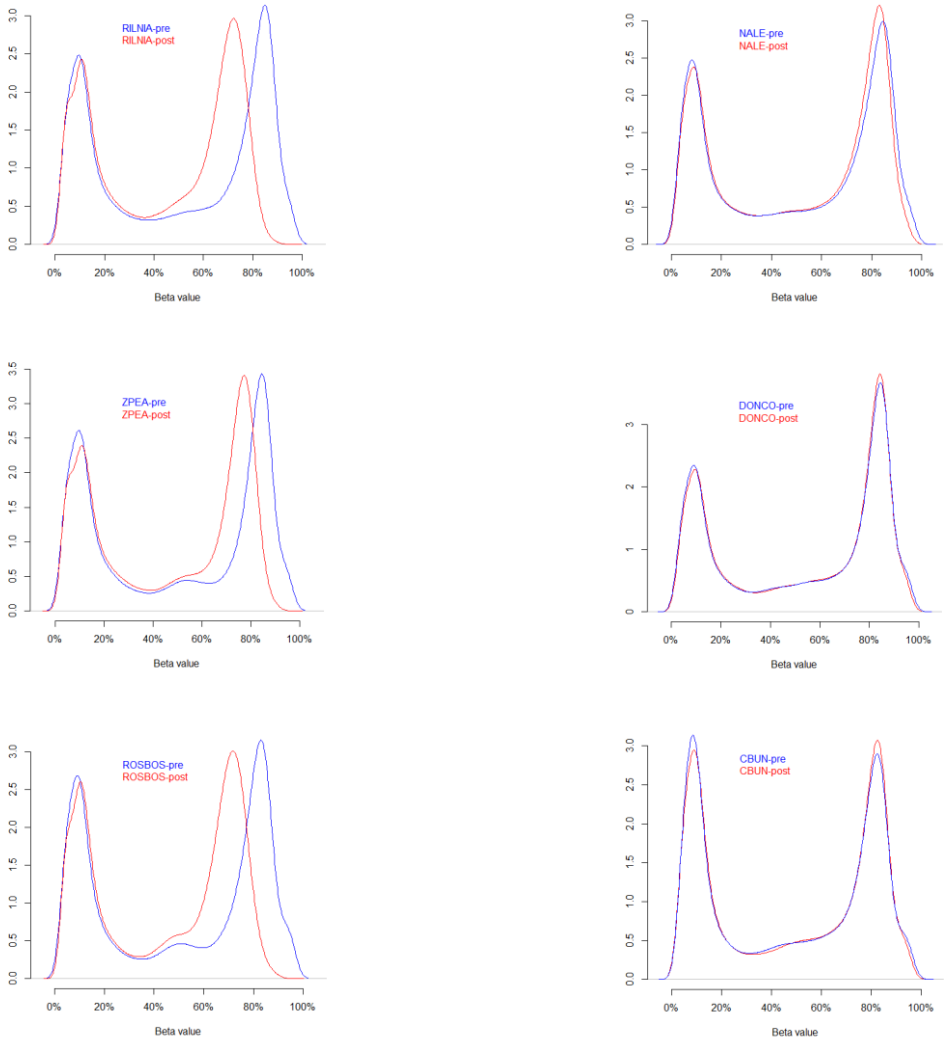


Figure 14 - kernel density plots of distributions of pre- and post-treatment beta values for each cell line and patient sample

For the two cell lines and three patient samples on the left hand side (RILNIA, ROSBOS and ZPEA), there is a clear difference between the pre- and post-treatment methylation profiles, whereas for the other five patient samples on the right there is little difference. For the former, there is a clear overall reduction in methylation levels following treatment, whereas for the latter group there is virtually no change.

This variation is also illustrated by figure 15 below, which shows the distribution of beta value changes across all CpG sites for each cell line and patient sample. For example, both cell lines have a small peak at around 30% demethylation, followed by a higher peak centred around 0% demethylation. RILNIA, ROSBOS and ZPEA have similar shapes, but with slightly higher peaks at around 10% rather than 30% demethylation. The other five patient samples have a single large peak centred around 0% demethylation.

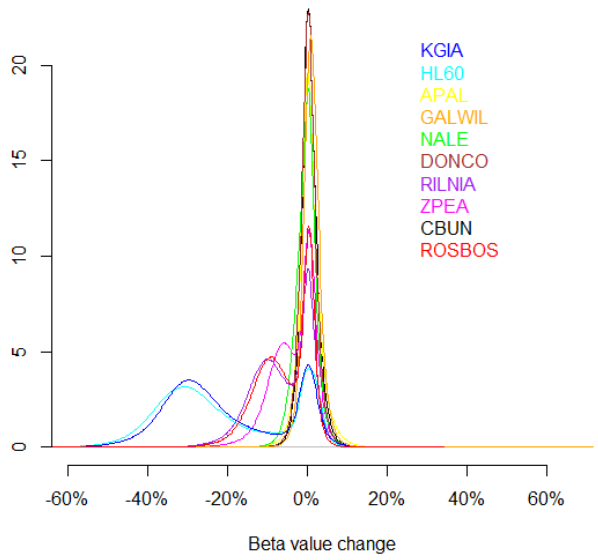


Figure 15 - kernel density plot of beta value changes after treatment for each cell line and patient sample

Table 4 below shows, for each cell line and patient sample, the proportions of CpG sites which fall into different bands of beta value change following treatment with DAC - for example, in the cell line HL60, 59.1% of CpG sites experienced an absolute reduction in beta value of at least 20% following treatment.

B-change	HL60	KGIA	APAL	GALWIL	NALE	DONCO	RILNIA	ZPEA	CBUN	ROSBOS
<-20%	59.1%	58.5%	0.0%	0.0%	0.0%	0.0%	3.1%	0.7%	0.0%	2.1%
-20 to -10%	10.4%	11.2%	0.2%	0.0%	0.3%	0.1%	27.7%	10.6%	0.0%	24.3%
-10 to 0%	15.0%	15.6%	43.9%	31.5%	53.9%	45.4%	45.5%	55.8%	41.6%	47.0%
0 to 10%	15.4%	14.4%	55.1%	68.2%	45.7%	54.4%	23.6%	32.8%	58.2%	26.5%
10 to 20%	0.1%	0.2%	0.8%	0.3%	0.2%	0.1%	0.2%	0.2%	0.2%	0.0%
>20%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%

Table 4 - distributions of beta value changes for each cell line and patient sample

The two cell lines, and to a lesser extent patient samples RILNIA, ROSBOS and ZPEA, (all five hereafter referred to as "group 1"), have significant proportions of CpG sites which experienced beta value reductions of at least 10% following treatment with DAC. On the other hand, for patient samples APAL, GALWIL, NALE, DONCO and CBUN ("group 2"), at least 99% of CpG sites experienced a change in beta value (in either direction) of less than 10%.

Summary of key points

The extent of demethylation following treatment with DAC varied between the two cell lines and eight patients samples. The two cell lines experienced the highest levels of demethylation, followed by patient samples RILNIA, ROSBOS and ZPEA. The other five patient samples experienced very little demethylation.

3.4 Variation in demethylation levels across gene locations and CpG island regions and with CPG density

The paragraphs below analyse how the changes in methylation level, for each cell line and patient sample, vary across gene locations, CpG island regions and with CpG density.

3.4.1 Average change in methylation level stratified by gene location

Figure 16 below shows the average beta values after treatment for each cell line and patient sample stratified by gene location.

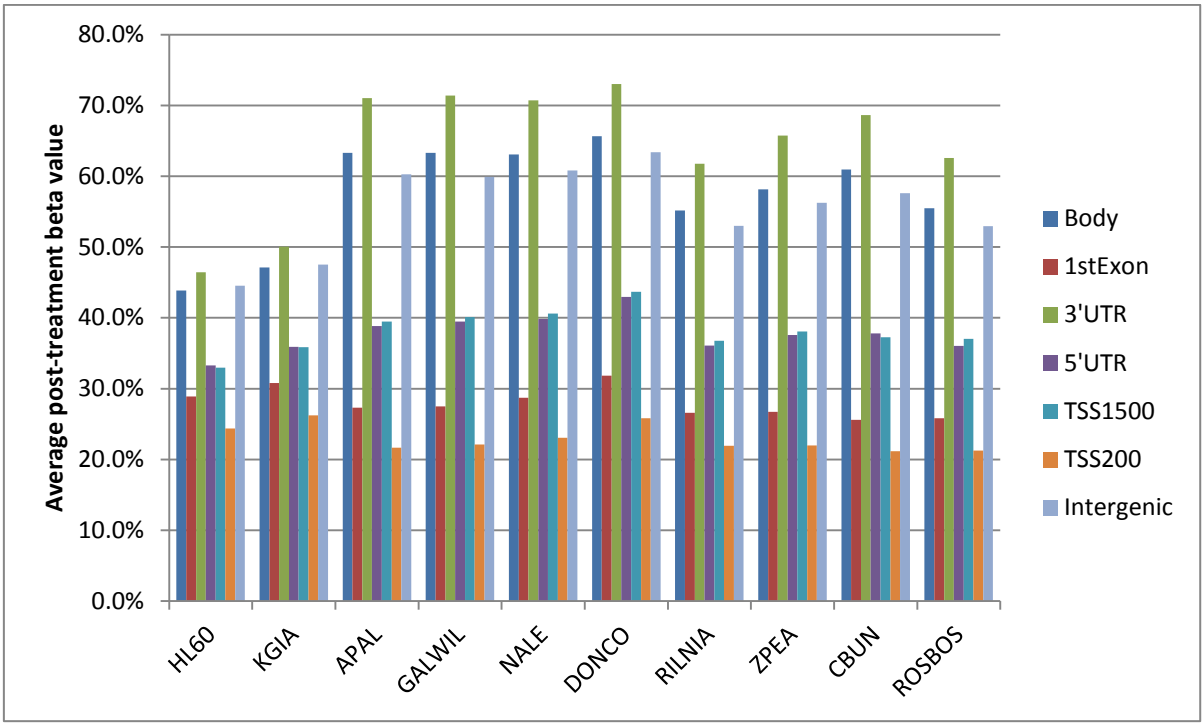


Figure 16 - average post-treatment beta values for each gene location

Comparison with figure 9 shows the extent of demethylation across gene locations. This is most apparent for the two cell-lines, followed by patient samples RILNIA, ROSBOS and ZPEA.

Figures 17 and 18 below show the average absolute and average proportionate beta value reductions (i.e. a positive number represents a reduction in beta value) respectively for each cell line and patient sample stratified by gene location.

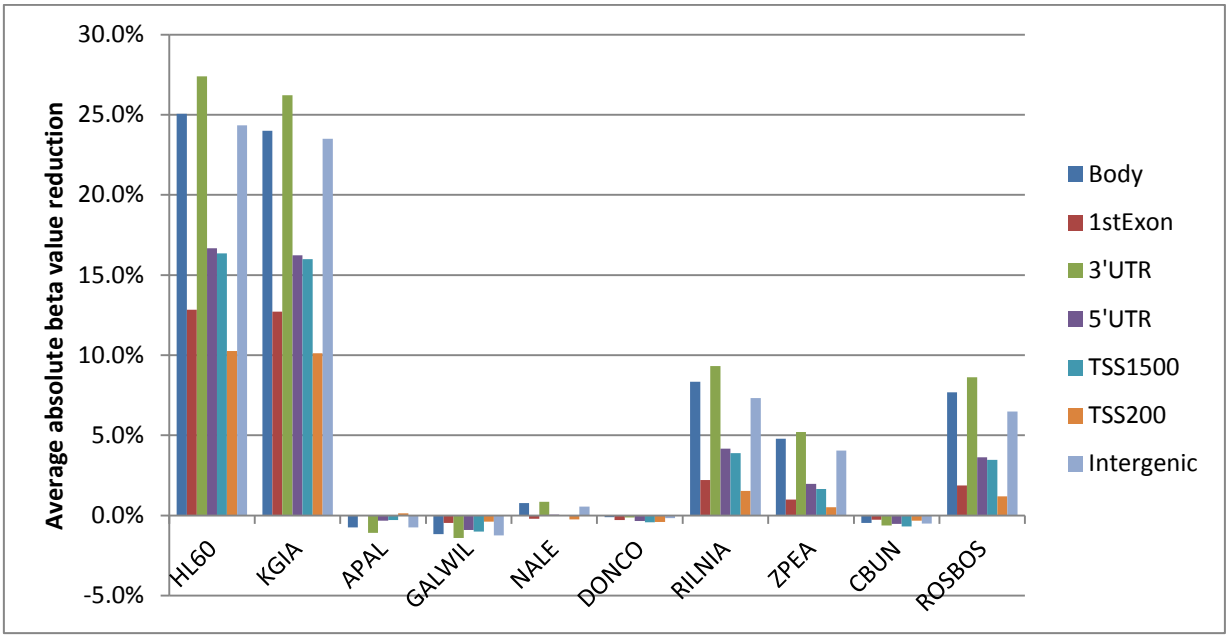


Figure 17 - average absolute beta value reductions for each gene location

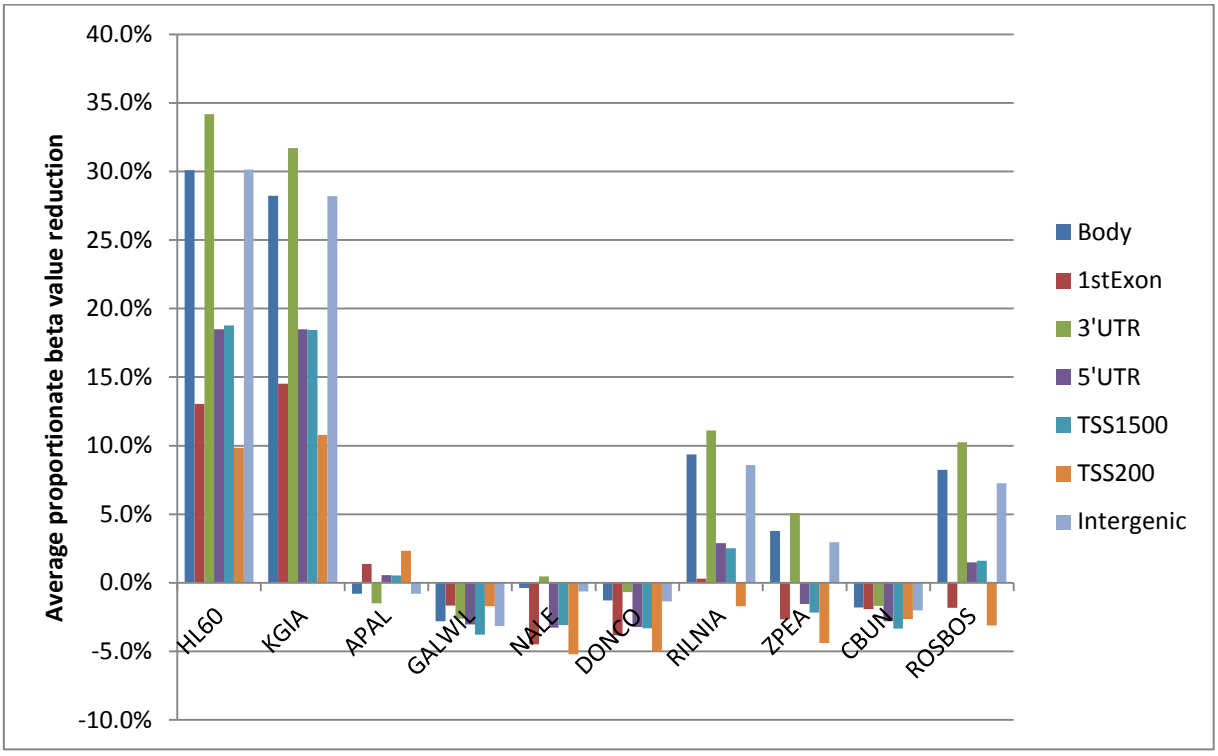


Figure 18 - average proportionate beta value reductions for each gene location

For all members of group 1, there was, on average, an absolute reduction in methylation level after treatment across all gene locations. CpG sites located in 3'UTRs experienced the largest average reductions in methylation, followed closely by sites in bodies and intergenic regions.

Sites located in TSS200s experienced the lowest average reductions. The pattern of reductions across gene locations was similar when proportionate changes were considered.

3.4.2 Average change in methylation level stratified by CpG island region

Figure 19 below shows the average beta values after treatment for each cell line and patient sample stratified by CpG island region.

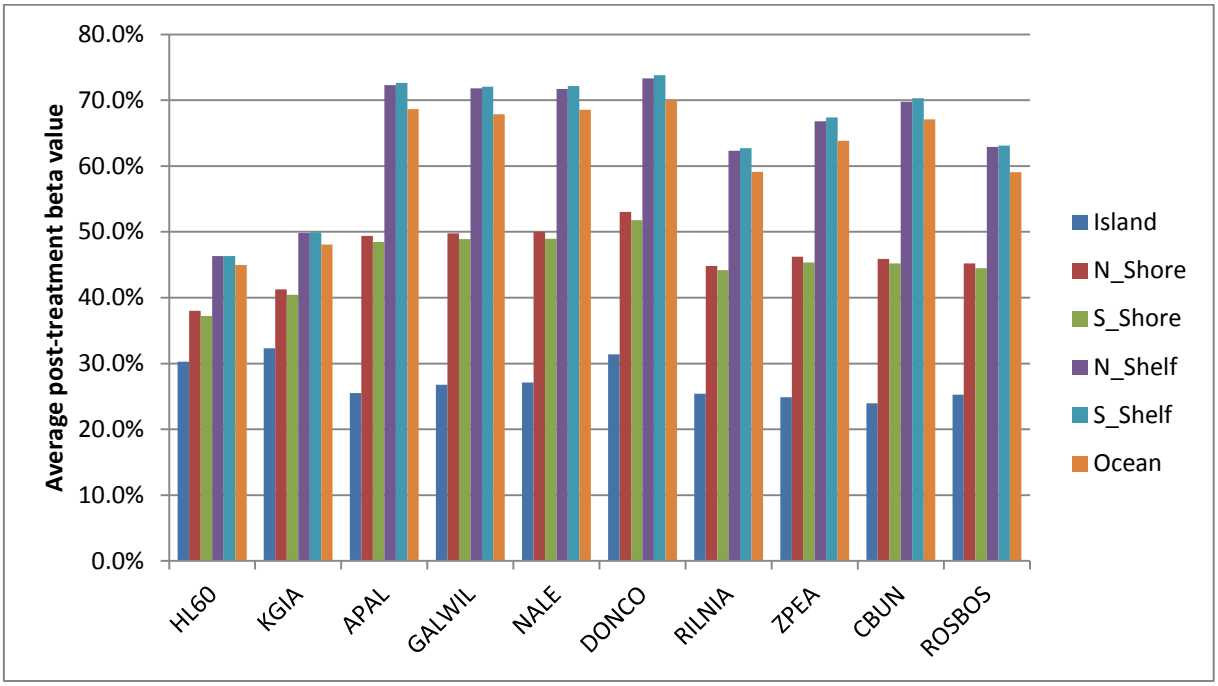


Figure 19 - average post-treatment beta values for each CpG island region

Comparison with figure 10 shows the extent of demethylation across CpG island regions. Again, this is most apparent for the two cell-lines, followed by patient samples RILNIA, ROSBOS and ZPEA.

Figures 20 and 21 below show the average absolute and average proportionate beta value reductions respectively for each cell line and patient sample stratified by CpG island region.

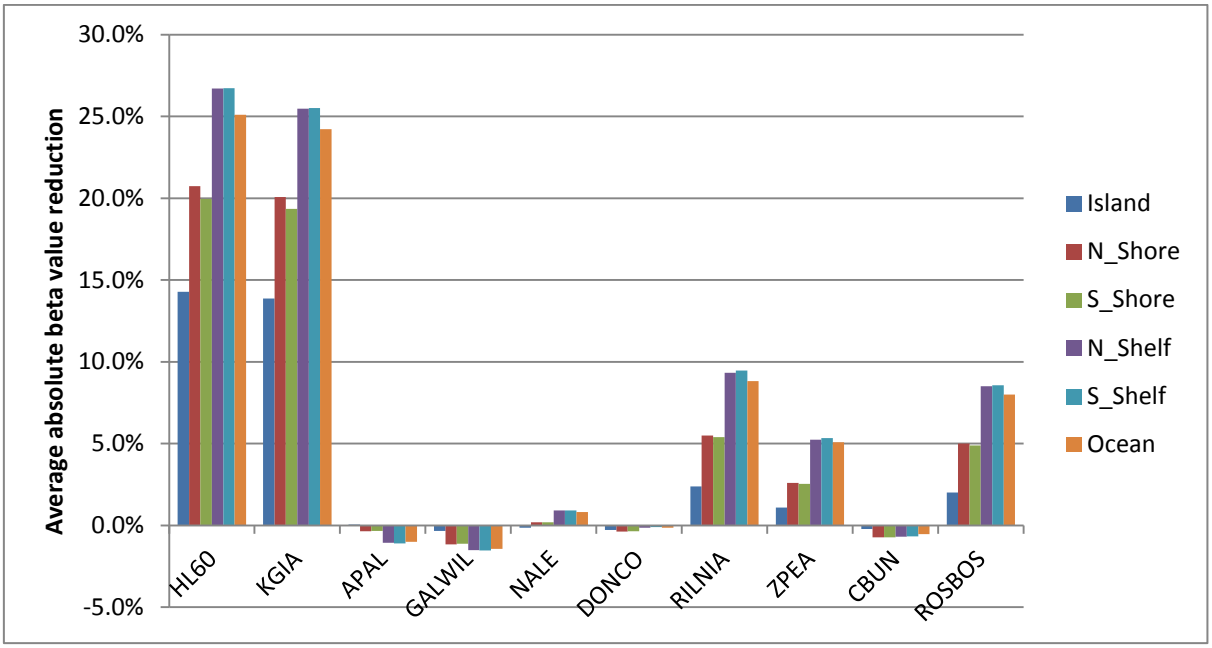


Figure 20 - average absolute beta value reductions for each CpG island region

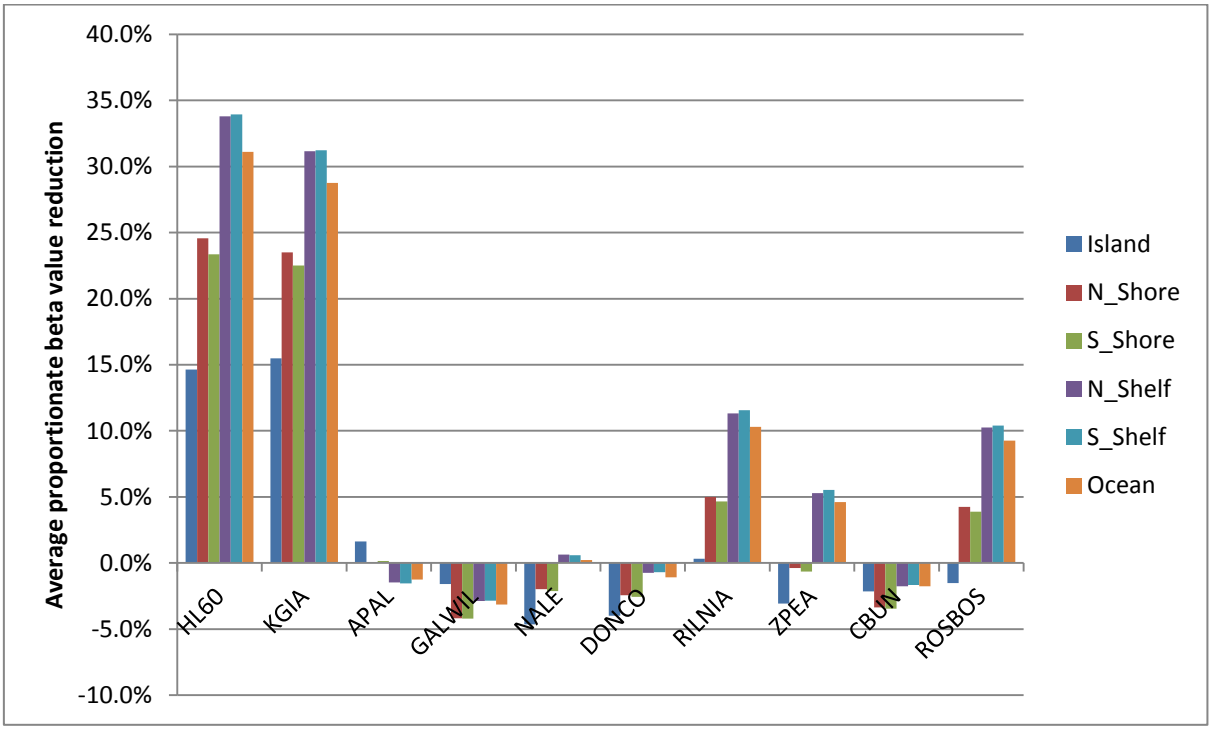


Figure 21 - average proportionate beta value reductions for each CpG island region

Again there is a consistent pattern amongst the members of group 1, and the pattern is broadly consistent using absolute or proportionate changes. This time, CpG sites located in north and south shelves experienced the largest average reductions in methylation, followed closely by

sites in ocean regions. Sites located in island regions had the lowest average methylation reductions.

3.4.3 Average change in methylation level stratified by CpG density

Figure 22 below shows the average beta values after treatment for each cell line and patient sample stratified by CpG density group.

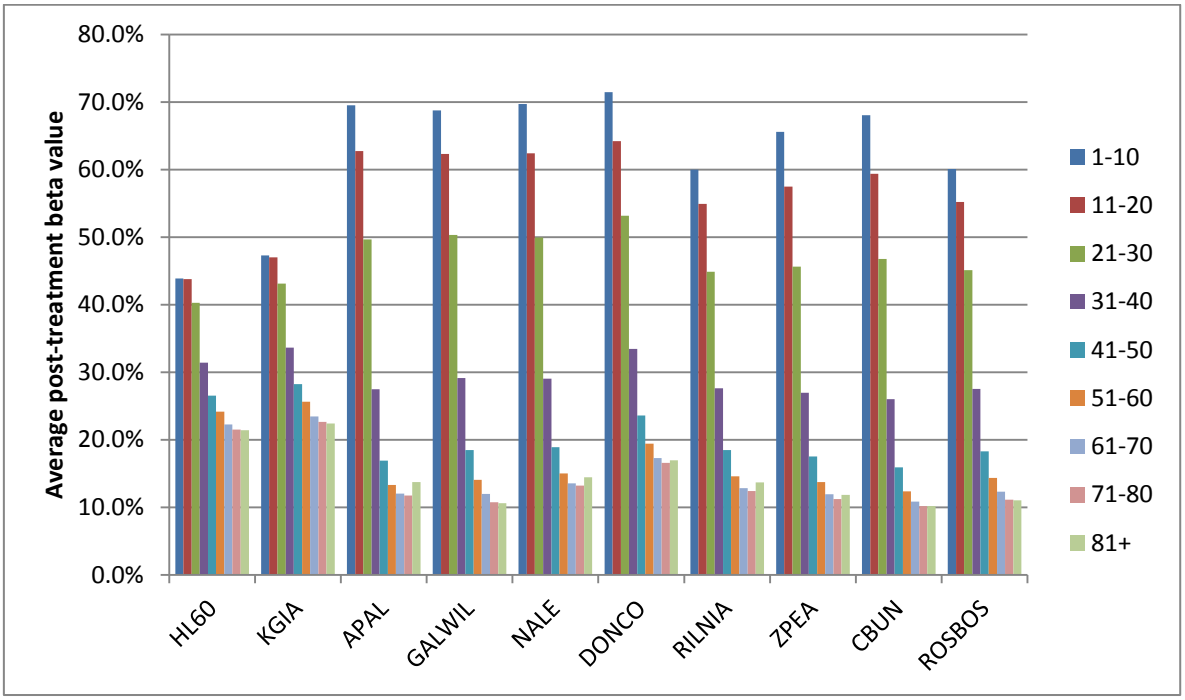


Figure 22 - average post-treatment beta values for each CpG density group

Comparison with figure 11 shows the extent of demethylation across CpG density groups. Once again, this is most apparent for the two cell-lines, followed by patient samples RILNIA, ROSBOS and ZPEA.

Figures 23 and 24 below show the average absolute and average proportionate beta value reductions respectively for each cell line and patient sample stratified by CpG density group.

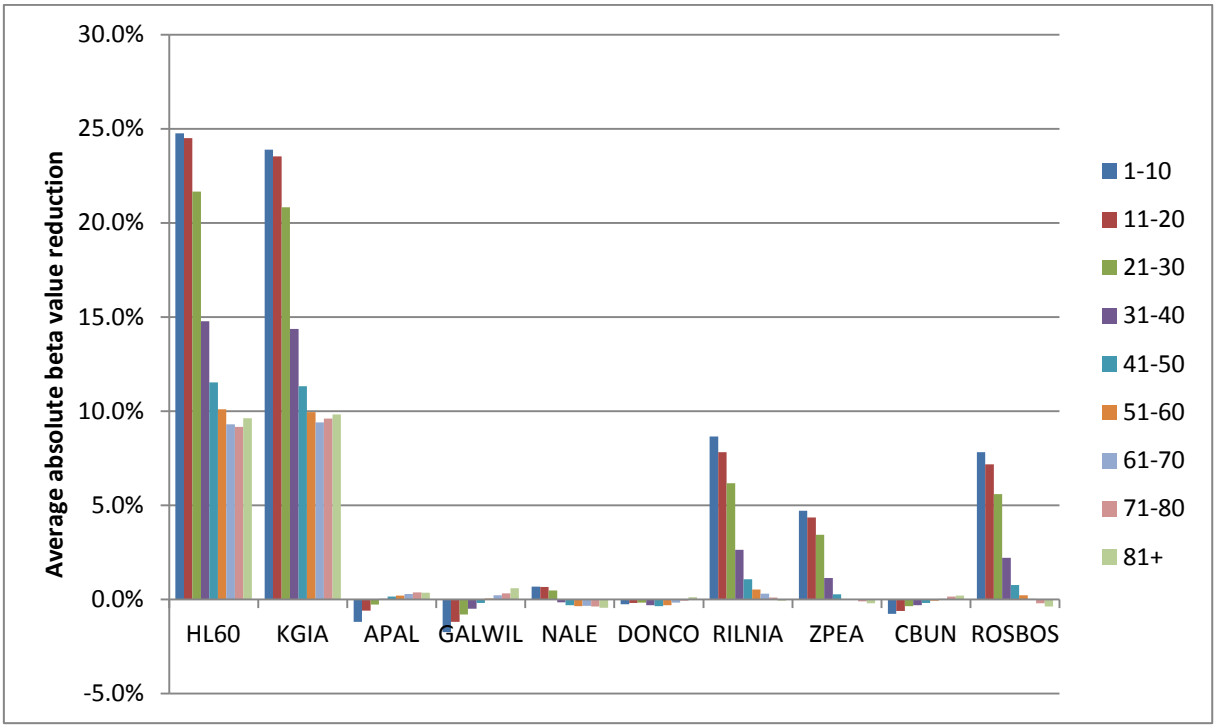


Figure 23 - average absolute beta value reductions for each CpG density group

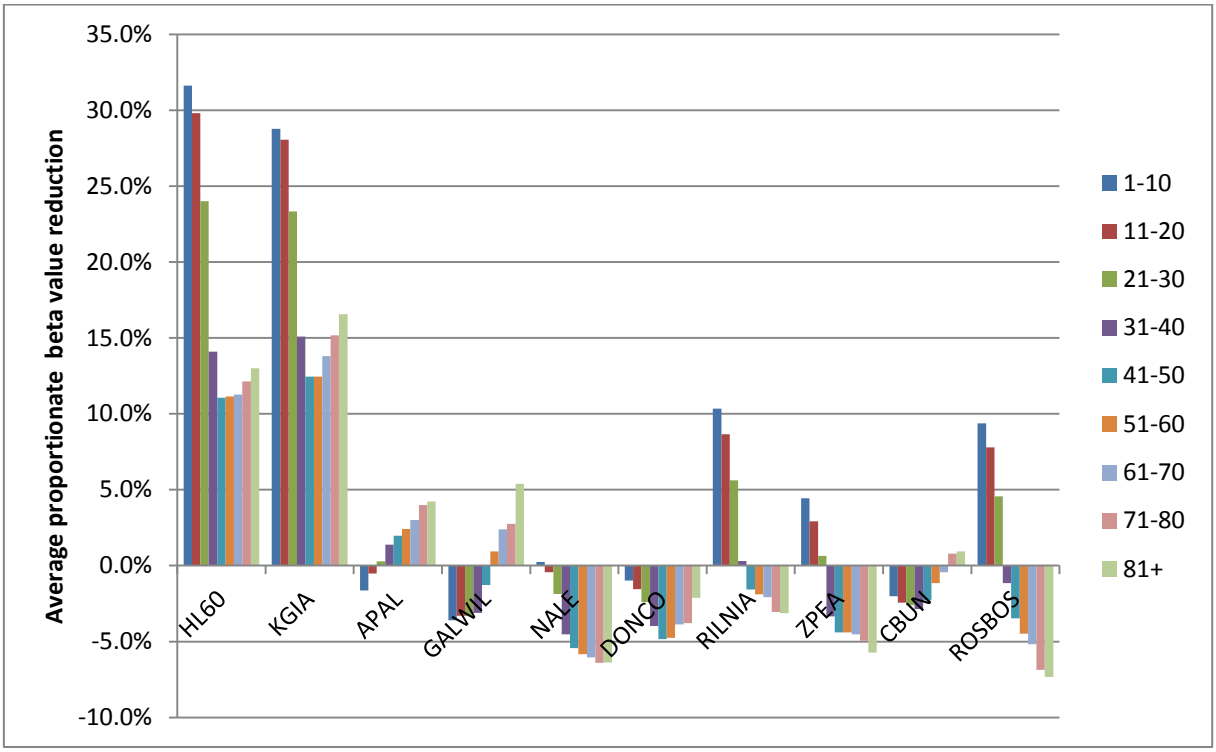


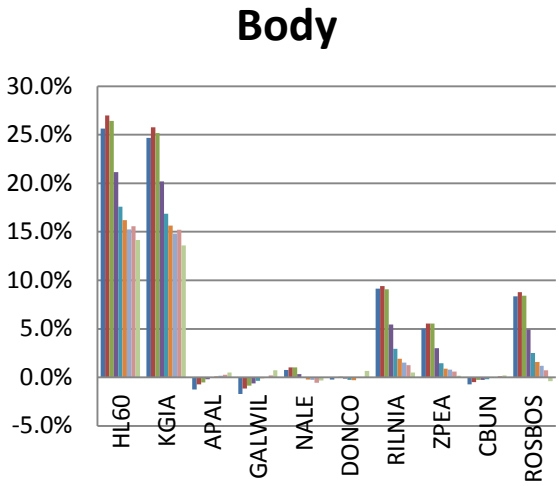
Figure 24 - average proportionate beta value reductions for each CpG density group

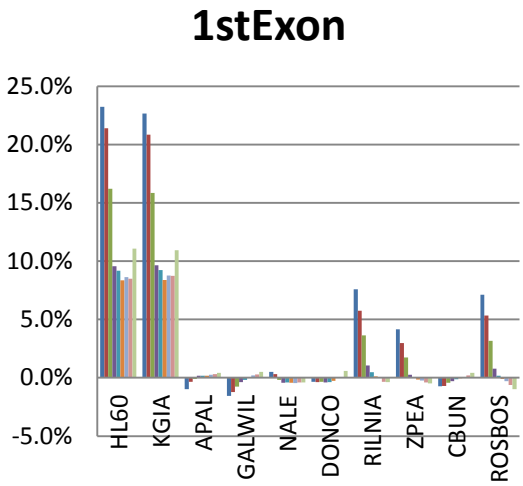
For all five members of group 1 there is a clear trend for the absolute reduction in methylation level to decrease as CpG density increases. However, for the two cell lines the trend flattens out at an absolute beta value reduction of just under 10% for the higher CpG densities (CpG count greater than 60), whereas it continues sharply downwards for RILNIA, ROSBOS and ZPEA. The pattern of proportionate changes is similar to that for absolute changes, although at higher CpG densities some small proportionate increases are observed in the patient samples. CpG sites with high CpG densities tend to have low pre-treatment beta values, so some care is required in interpreting these proportionate changes.

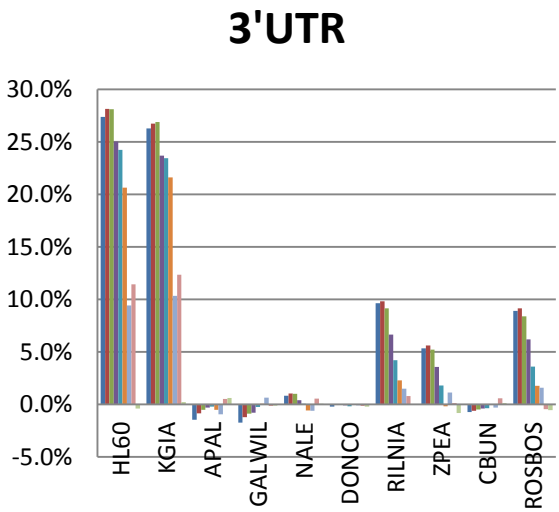
The next two paragraphs consider how the average reductions in methylation level of CpG sites in each gene location and island region vary across areas of different CpG density.

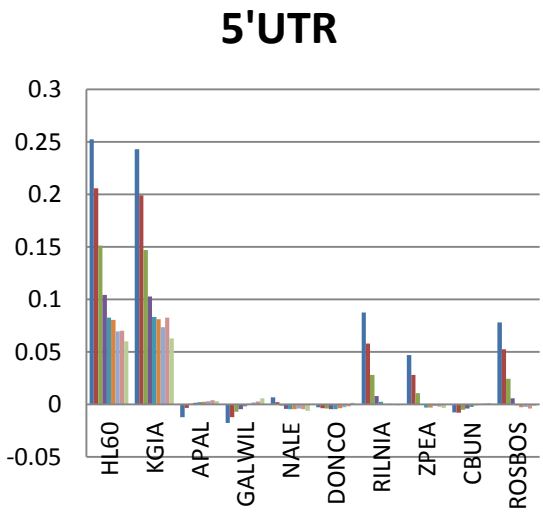
3.4.4 Average change in methylation level for each gene location stratified by CpG density

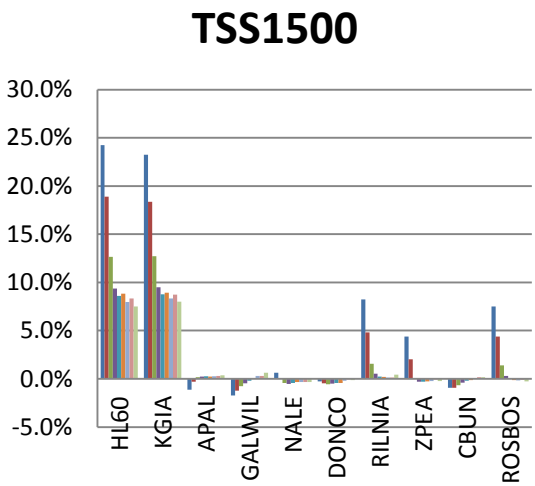
Figure 25 below shows, for each cell line and patient sample, the average beta value reductions (y-axis) for each gene location, stratified by CpG density group. The trend for decreasing average methylation reductions as CpG density increases is present across all gene locations. However, it is less pronounced in bodies, 3'UTRs and intergenic regions. Also, as previously noted, the trend flattens out in the two cell lines at higher CpG densities across all gene locations.

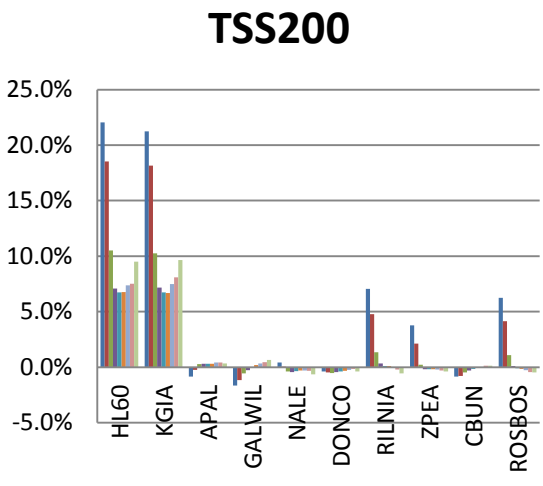

Body
30.0%
25.0%
20.0%
15.0%
10.0%
5.0%
0.0%
-5.0%
HL60
KGIA
APAL
GALWIL
NALE
DONCO
RILNIA
ZPEA
CBUN
ROSBOS


1stExon
25.0%
20.0%
15.0%
10.0%
5.0%
0.0%
-5.0%
HL60
KGIA
APAL
GALWIL
NALE
DONCO
RILNIA
ZPEA
CBUN
ROSBOS


3'UTR
30.0%
25.0%
20.0%
15.0%
10.0%
5.0%
0.0%
-5.0%
HL60
KGIA
APAL
GALWIL
NALE
DONCO
RILNIA
ZPEA
CBUN
ROSBOS


5'UTR
0.3
0.25
0.2
0.15
0.1
0.05
0
-0.05
HL60
KGIA
APAL
GALWIL
NALE
DONCO
RILNIA
ZPEA
CBUN
ROSBOS


TSS1500
30.0%
25.0%
20.0%
15.0%
10.0%
5.0%
0.0%
-5.0%
HL60
KGIA
APAL
GALWIL
NALE
DONCO
RILNIA
ZPEA
CBUN
ROSBOS


TSS200
25.0%
20.0%
15.0%
10.0%
5.0%
0.0%
-5.0%
HL60
KGIA
APAL
GALWIL
NALE
DONCO
RILNIA
ZPEA
CBUN
ROSBOS

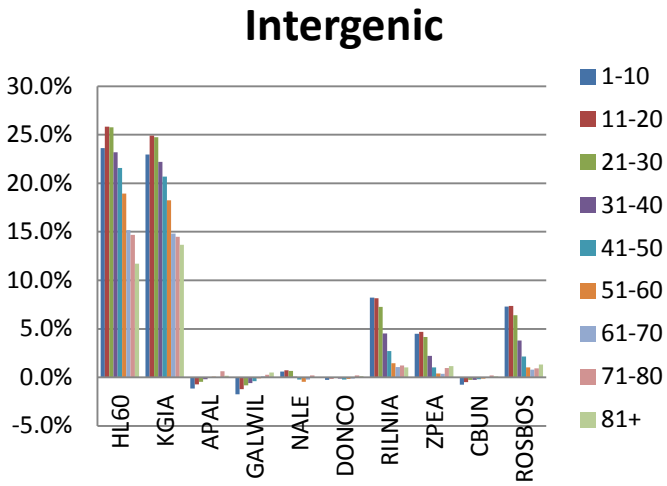
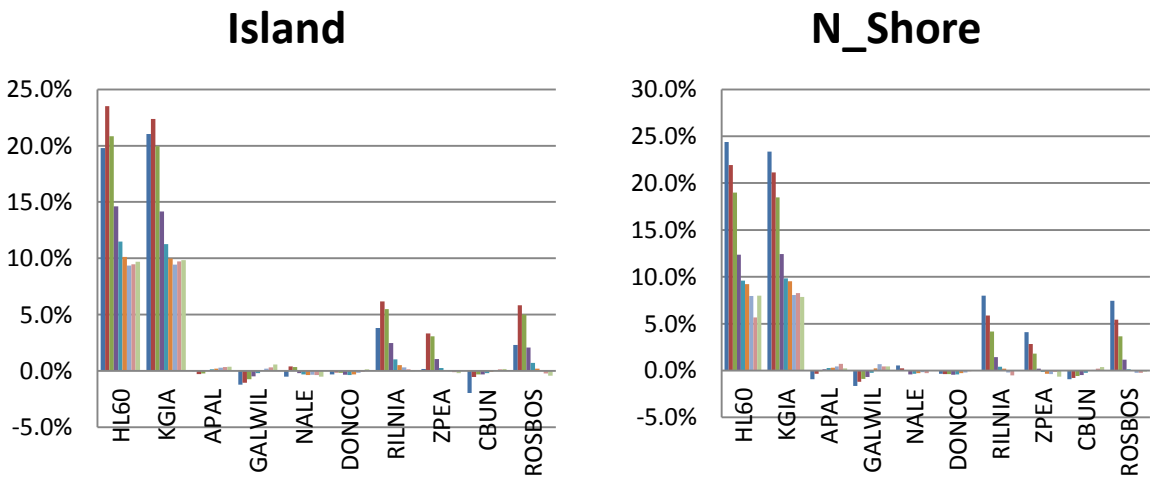


Figure 25 - average beta value reductions for each gene location, stratified by CpG density group

3.4.5 Average change in methylation level for each CpG island region stratified by CpG density

Figure 26 below shows, for each cell line and patient sample, the average beta value reductions (y-axis) for each CpG island region, stratified by CpG density group. The general trend for decreasing average methylation reductions as CpG density increases is again present across all CpG island regions. However, this trend is not entirely consistent, for example at higher CpG densities in both north and south shelves.



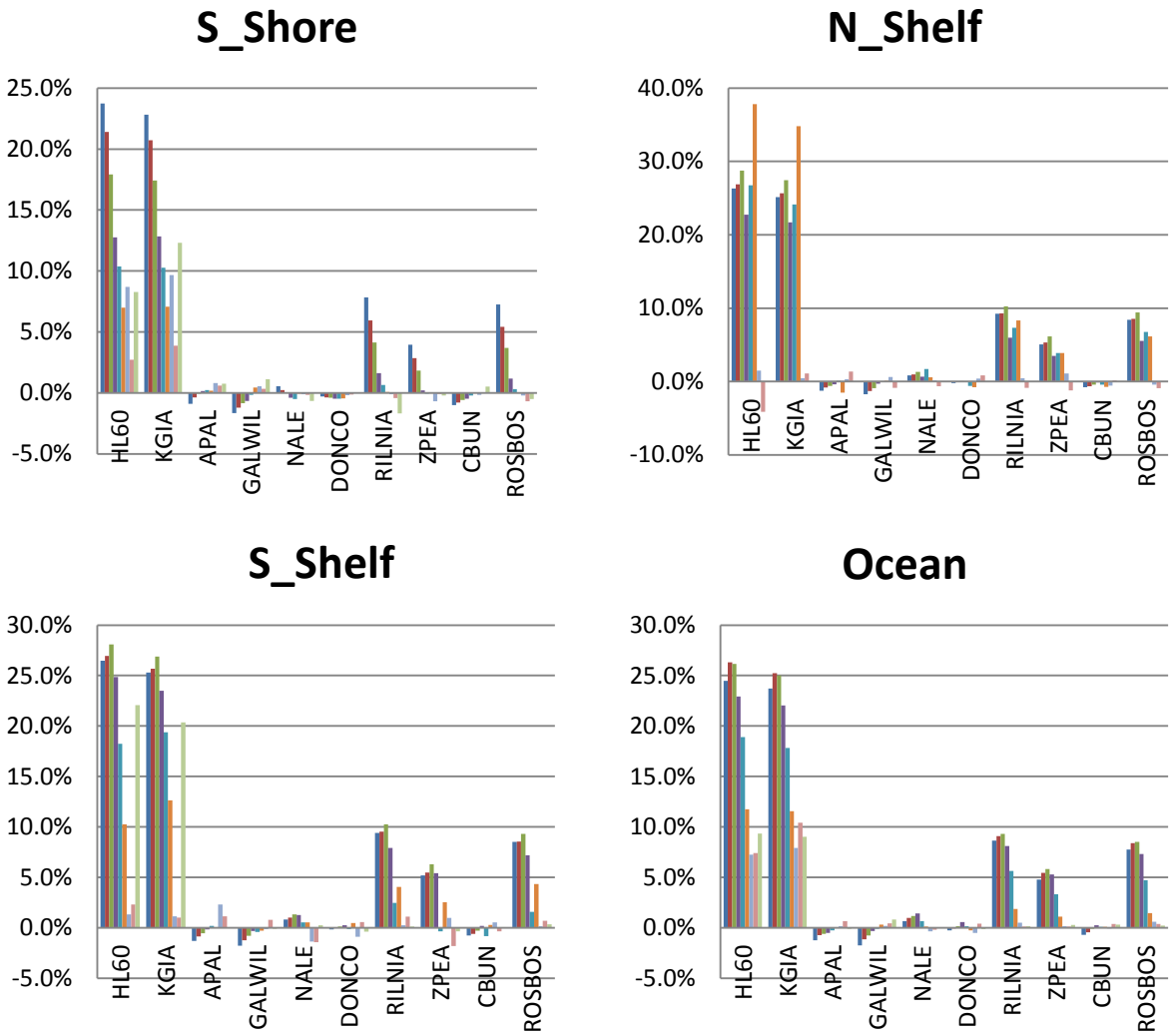


Figure 26 - average beta value reductions for each CpG island region, stratified by CpG density group (legend as Figure 25)

3.4.6 Summary of key points

The analysis above shows that the level of demethylation varies across gene locations and CpG island regions and with CpG density. In particular, it appears to be inversely correlated to CpG density, and this pattern is also seen when the analysis is broken down by both gene location and island region. Analyses by both absolute and proportionate changes show similar patterns of demethylation.

3.5 Variation in changes in methylation levels following treatment with DAC across pre-treatment methylation levels

This section considers how average reductions in methylation level vary across CpG sites according to pre-treatment methylation level, with further stratification by gene location, CpG island region and CpG density.

3.5.1 Average change in methylation level stratified by pre-treatment methylation level

Figure 27 below shows the average beta values after treatment for each cell line and patient sample stratified by pre-treatment beta value group.

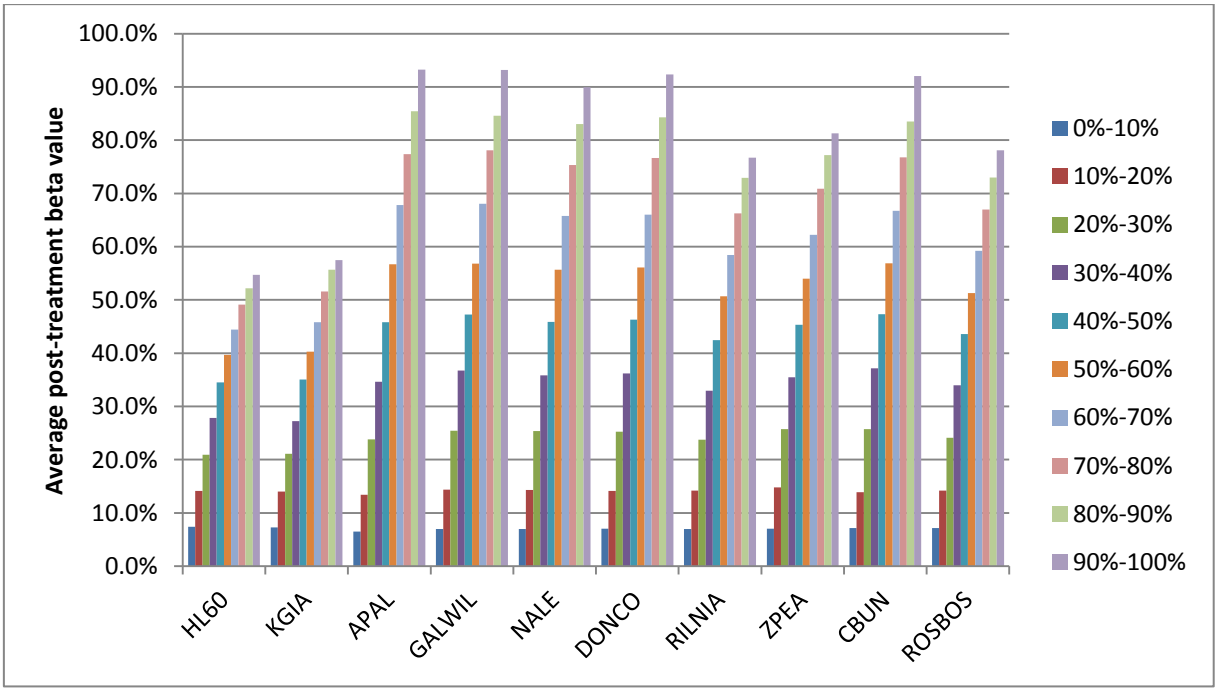


Figure 27 - average post-treatment beta values for each pre-treatment beta value group

The extent of demethylation across pre-treatment beta value groups is most apparent for the two cell-lines, followed by patient samples RILNIA, ROSBOS and ZPEA.

Figures 28 and 29 below show the average absolute and average proportionate beta value reductions respectively for each cell line and patient sample stratified by pre-treatment beta value group.

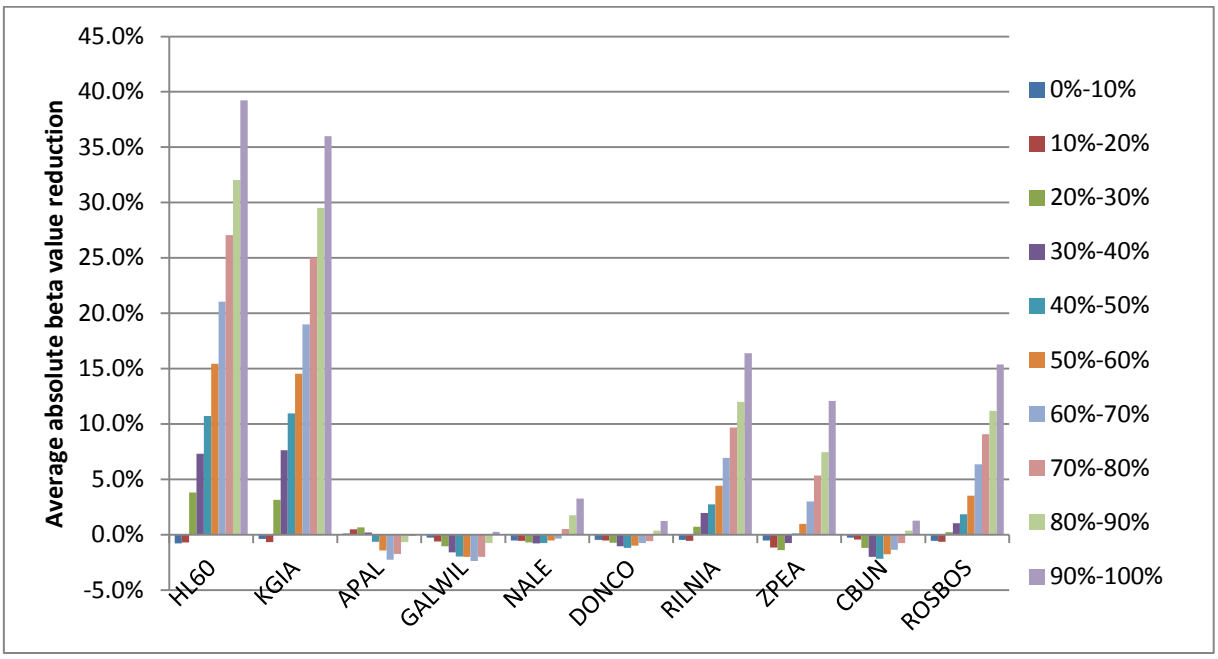


Figure 28 - average absolute beta value reductions for each pre-treatment beta value group

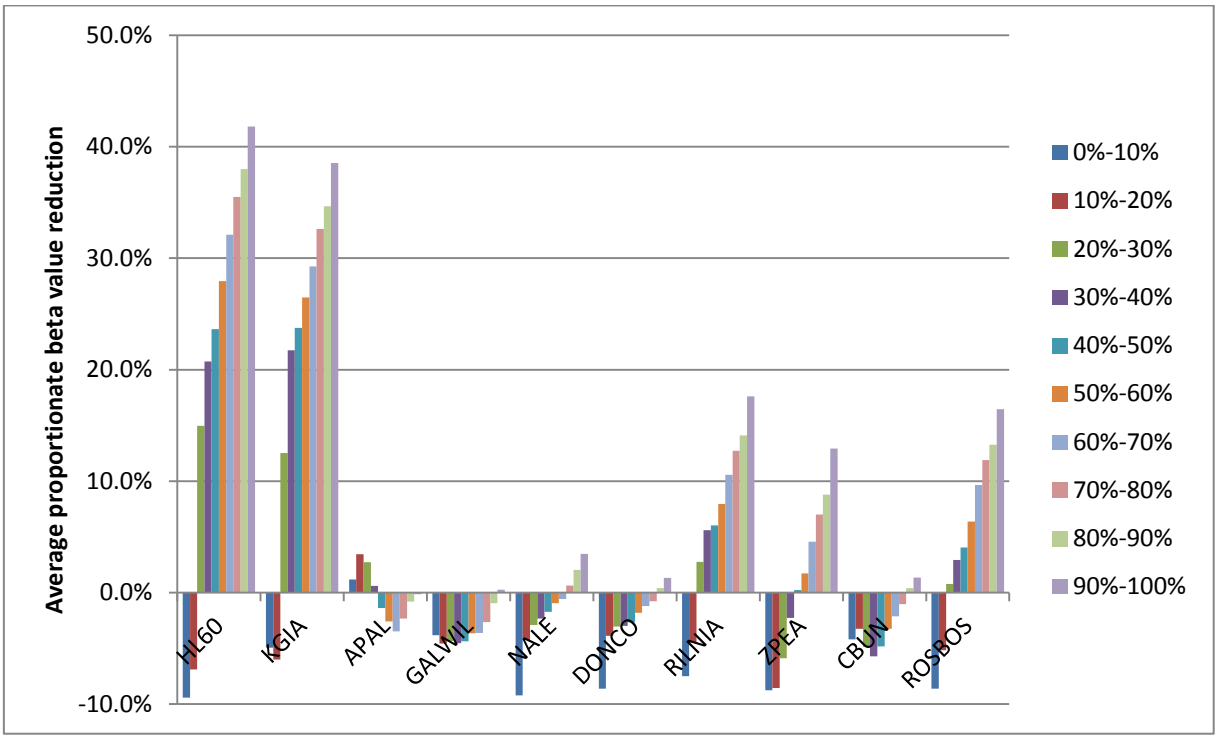


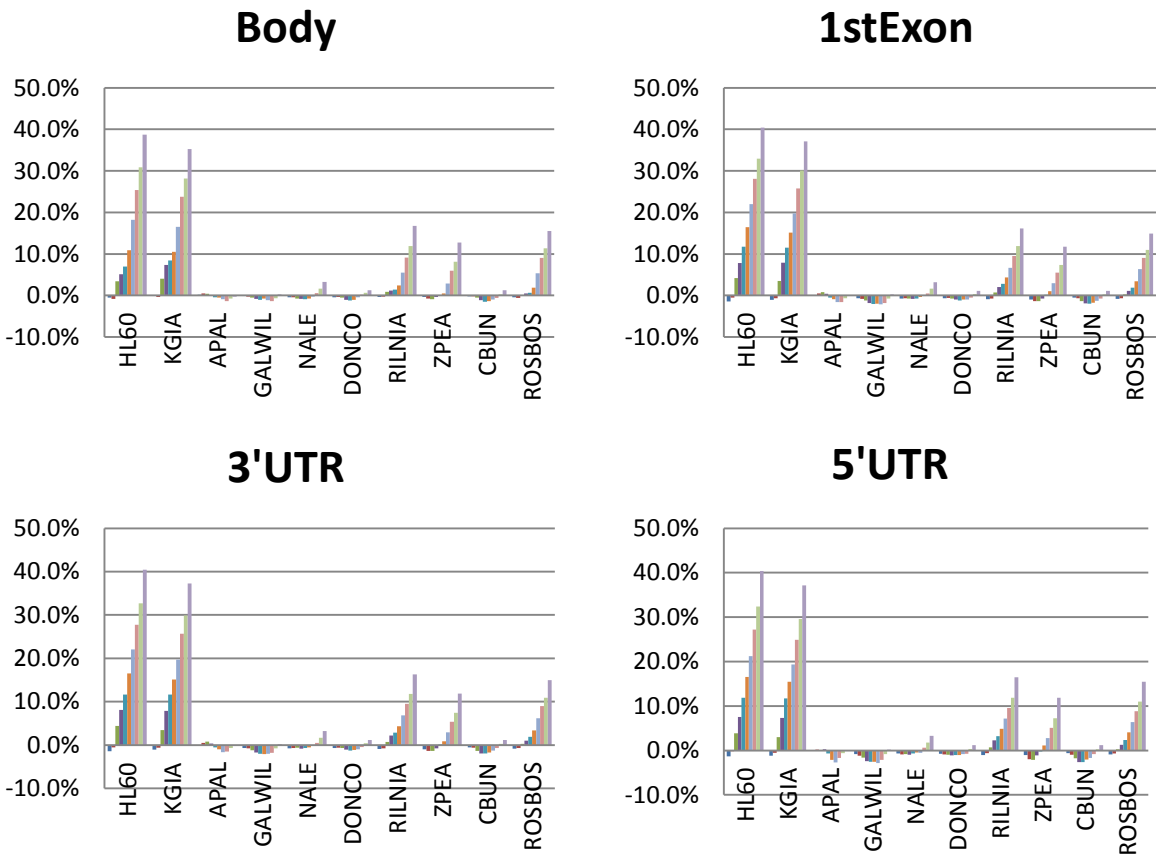
Figure 29 - average proportionate beta value reductions for each pre-treatment beta value group

Across all five members of group 1, there is a very clear trend of increasing reduction in methylation level as pre-treatment methylation level increases. This is observed for both absolute and proportionate changes, although the pattern is less pronounced for the latter.

3.5.2 Average change in methylation level for each gene location stratified by pre-treatment methylation levels

Figure 30 below shows, for each cell line and patient sample, the average beta value reductions (y-axis) for each gene location, stratified by pre-treatment beta value group.

For all five members of group 1 the trend for increasing demethylation as pre-treatment methylation level increases remains clear, and this pattern is similar across all the different gene locations.



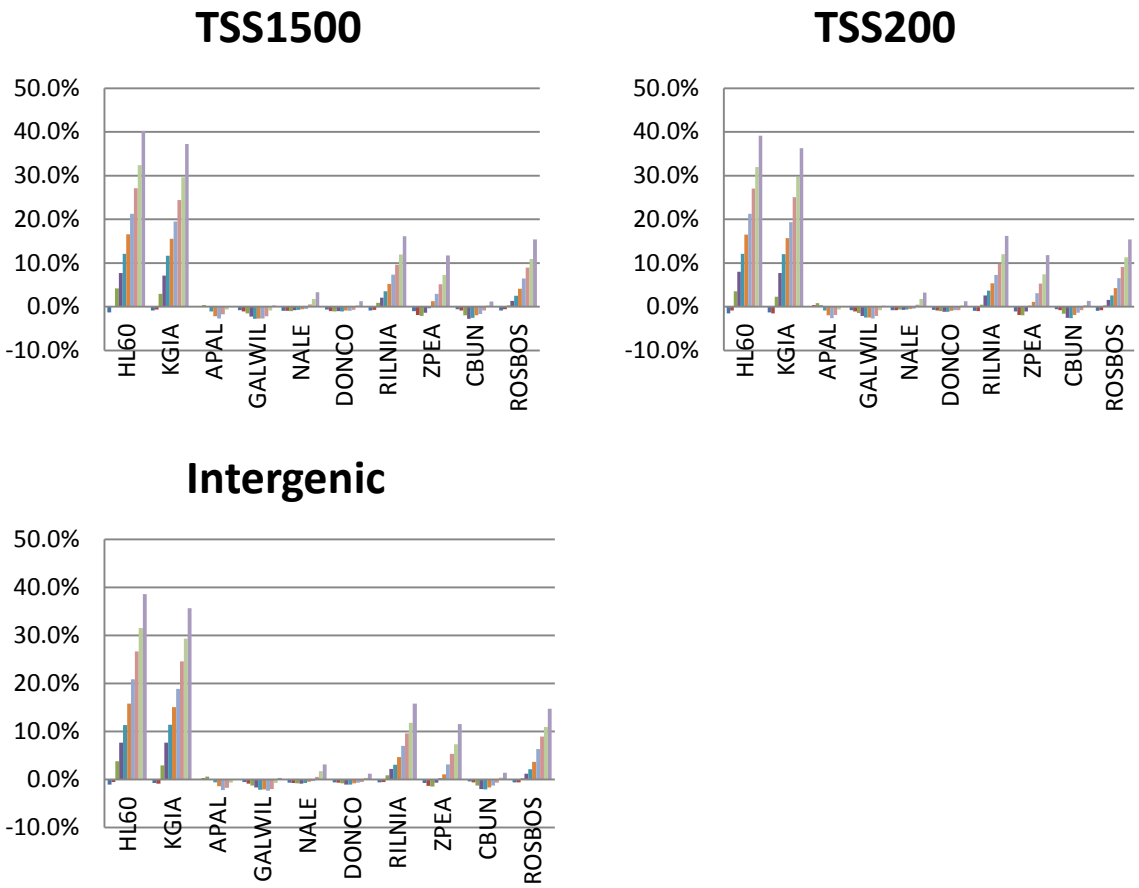


Figure 30 - average beta value reductions for each gene location, stratified by pre-treatment beta value group (legend as Figure 27)

3.5.3 Average change in methylation level for each CpG island region stratified by pre-treatment methylation levels

Figure 31 below shows, for each cell line and patient sample, the average beta value reductions (y-axis) for each CpG island region, stratified by pre-treatment beta value group. For all five members of group 1 the pattern of increasing demethylation with increasing pre-treatment methylation level is again clear, and this pattern is similar across all the different island region types.

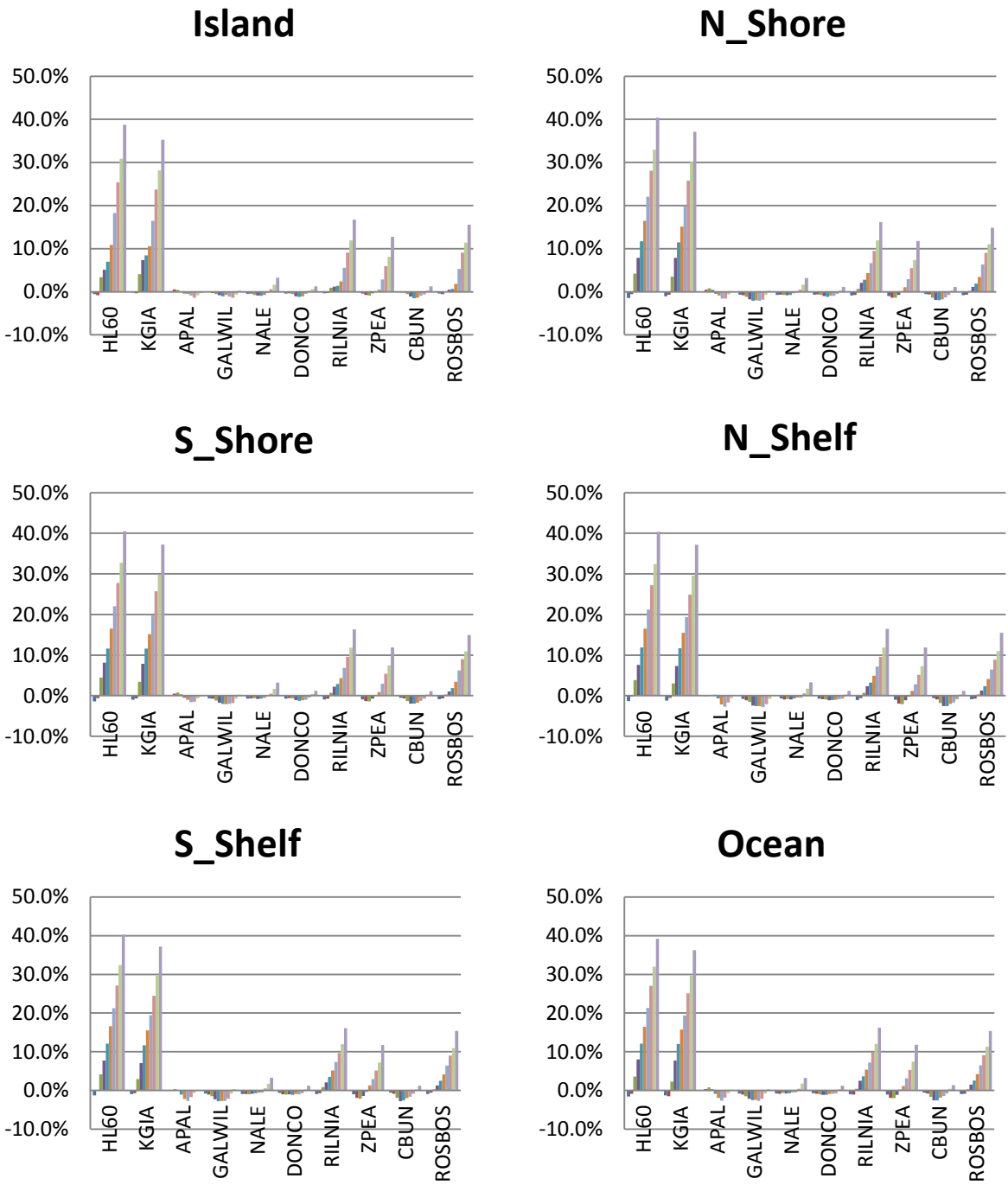
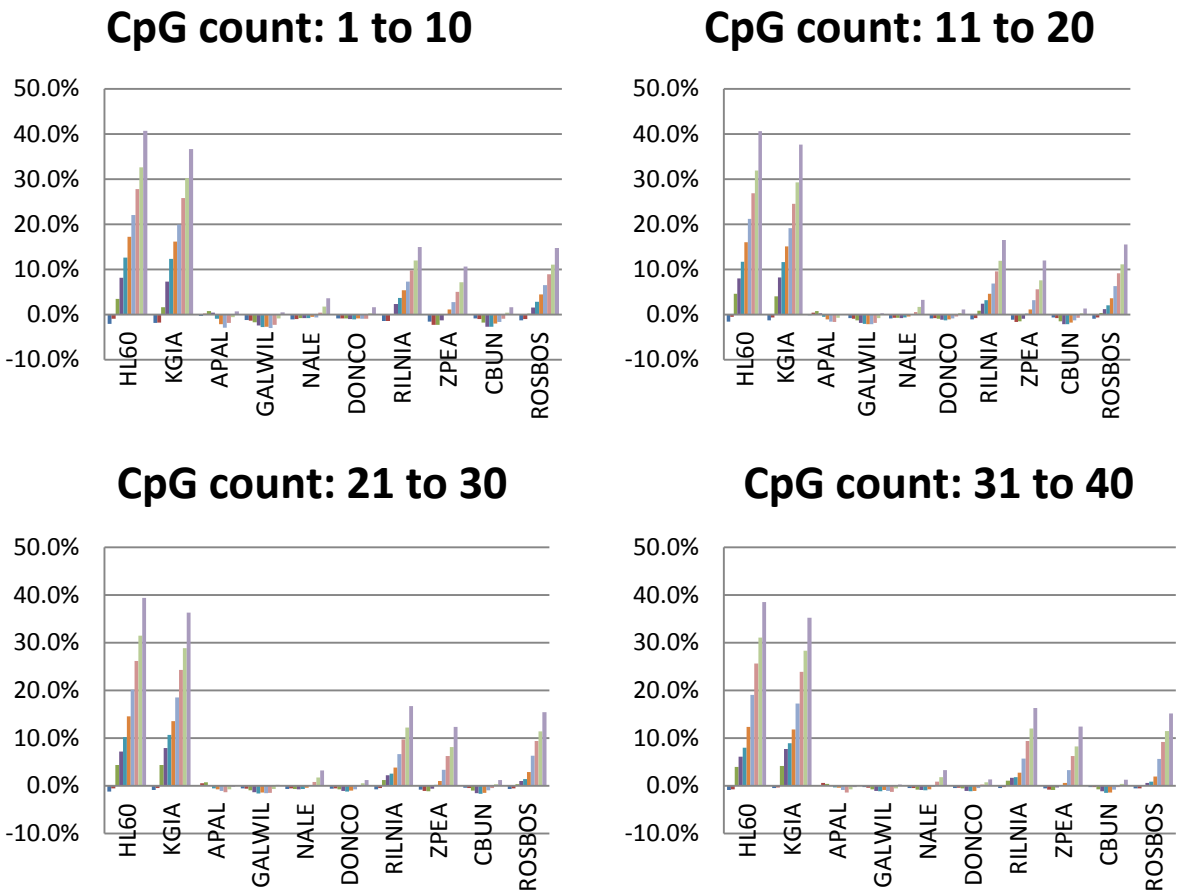


Figure 31 - average beta value reductions for each CpG island region, stratified by pre-treatment beta value group (legend as Figure 27)

3.5.4 Average change in methylation level for each CpG density group stratified by pre-treatment methylation levels

Figure 32 below shows, for each cell line and patient sample, the average beta value reductions (y-axis) for each CpG density group, stratified by pre-treatment beta value group. For both cell lines the trend for increasing demethylation with increasing pre-treatment methylation level is again clear. This is also true for the other three members of group 1, but the pattern does appear to break down at the highest CpG densities. This is particularly the case for ROSBOS in the highest CpG density category, where the average beta value reduction is around 25% for the 80-90% pre-treatment methylation group, but only around 15% for the 90-100% group.



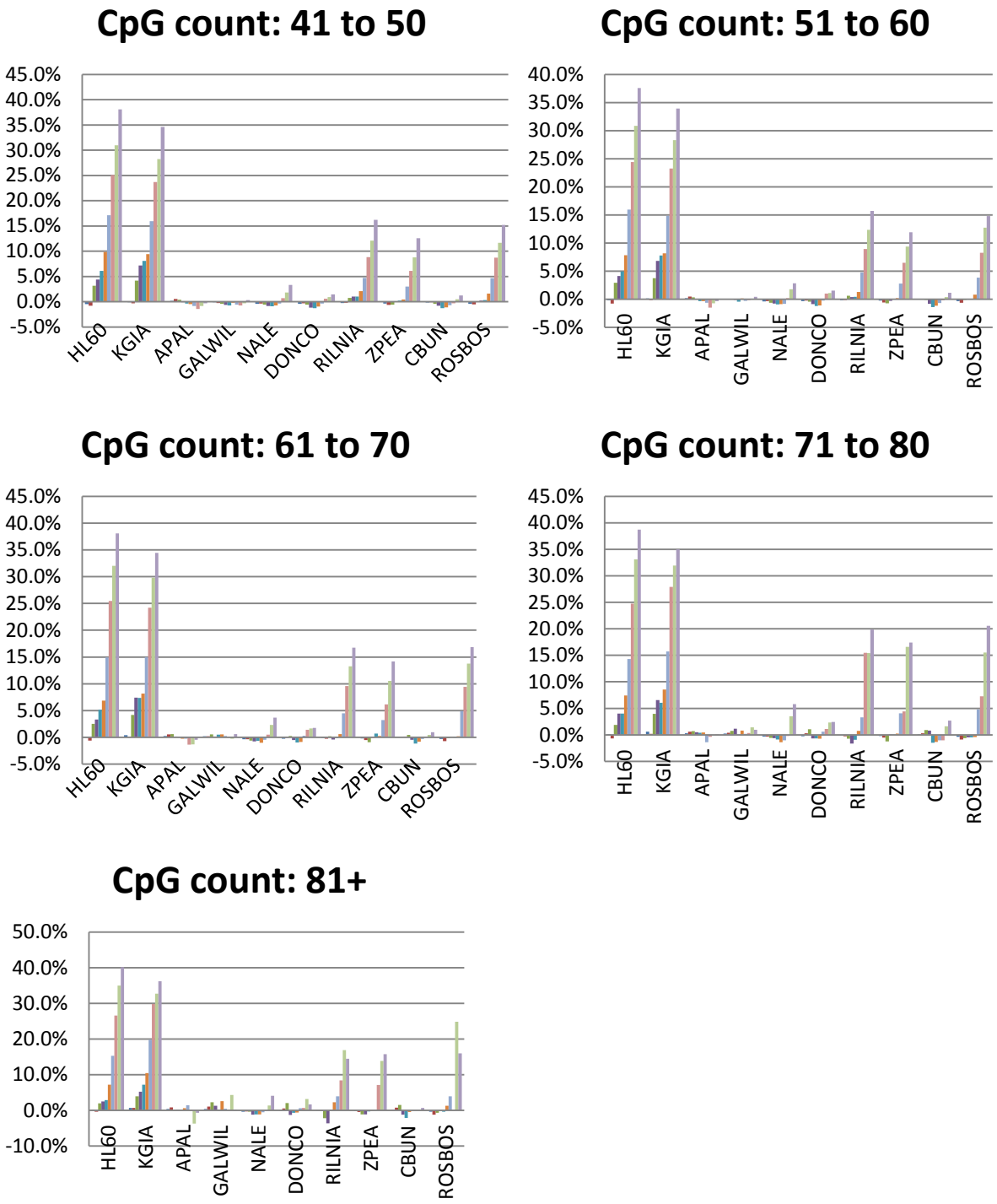


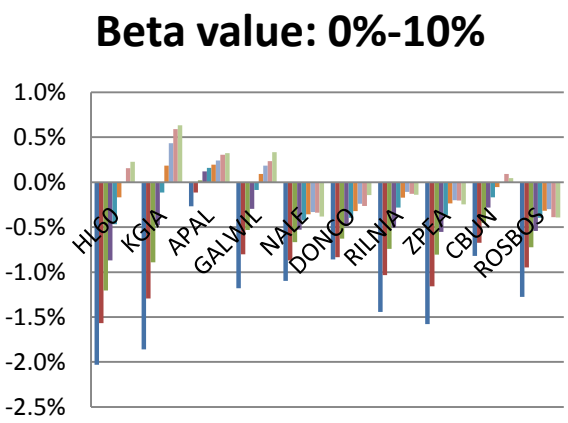
Figure 32 - average beta value reductions for each CpG density group, stratified by pre-treatment beta value group (legend as Figure 27)

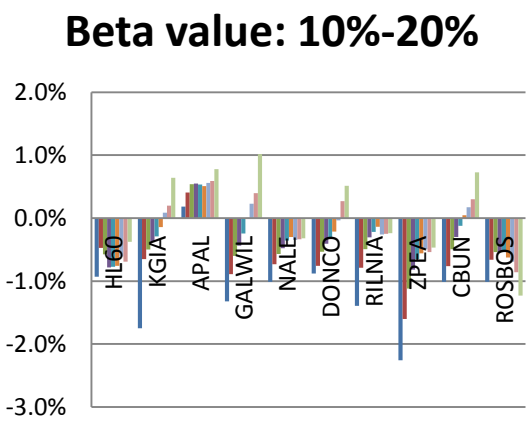
3.5.5 Conclusion so far

The analysis in this section has so far shown that pre-treatment methylation level appears to be the dominant factor in determining the extent to which a CpG site will be demethylated following treatment with DAC. However, the analysis in paragraph 3.5.4 suggests that there may also be a secondary effect associated with CpG density. In order to investigate this possibility further, paragraph 3.5.6 considers the analysis in 3.5.4 in the reverse order of stratification.

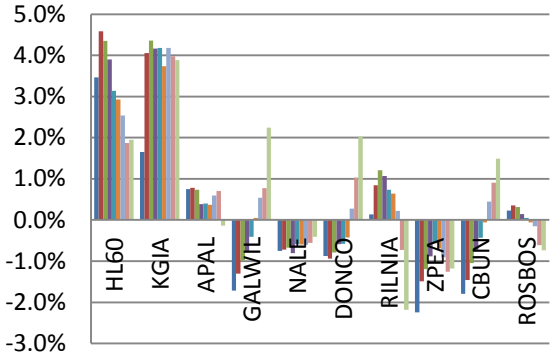
3.5.6 Average change in methylation level for each pre-treatment beta value group stratified by CpG density group

Figure 33 below shows, for each cell line and patient sample, the average beta value reductions (y-axis) for each pre-treatment beta value group, stratified by CpG density group. When the stratification is performed this way round, the variation across CpG density groups for any given pre-treatment beta value group is small. However, there is still some variation, in particular for CpG sites in the 40%-60% pre-treatment beta value range. Hence, whilst pre-treatment methylation level appears to have the most dominant association with demethylation, it is also possible that there is a secondary association with CpG density.

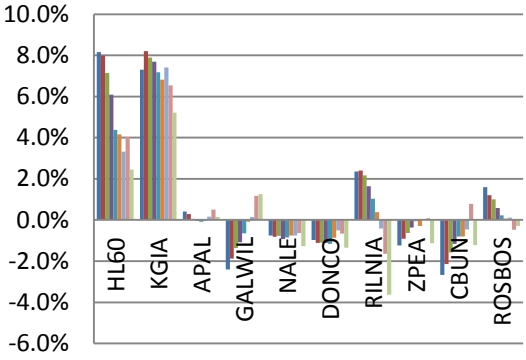




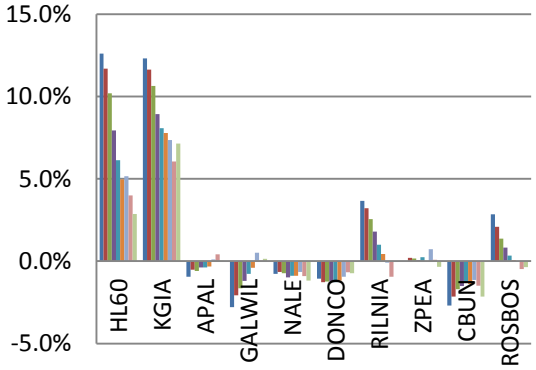
Beta value: 20%-30%



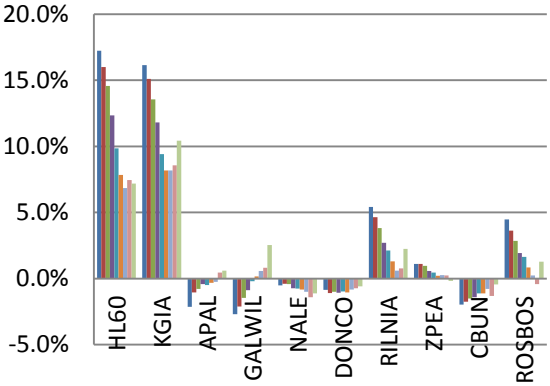
Beta value: 30%-40%



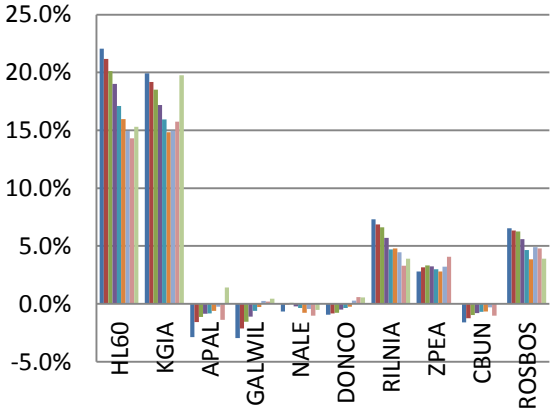
Beta value: 40%-50%



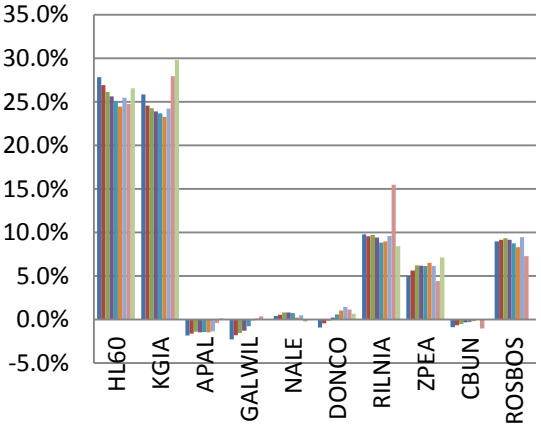
Beta value: 50%-60%



Beta value: 60%-70%



Beta value: 70%-80%



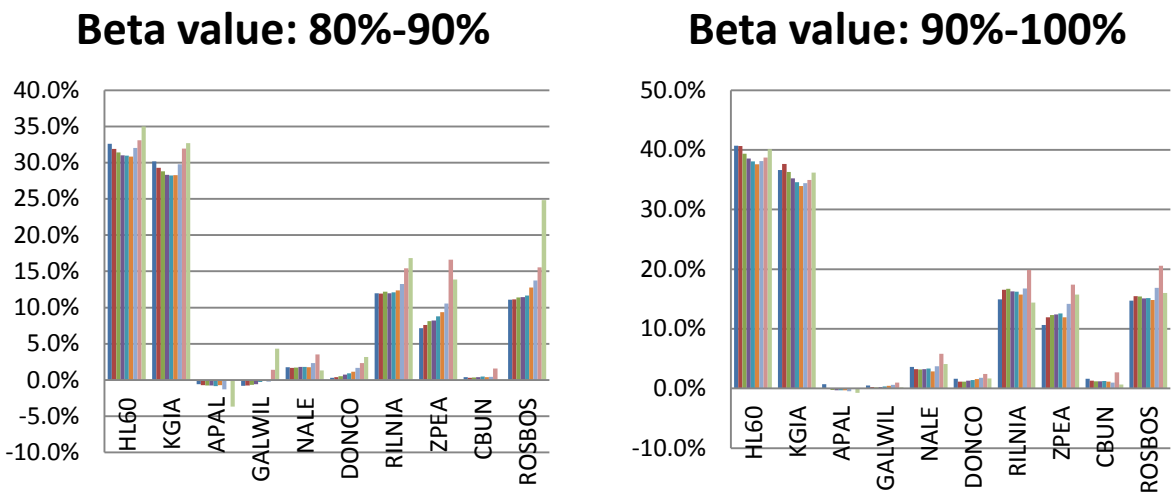


Figure 33 - average beta value reductions for each pre-treatment beta value group, stratified by CpG density group (legend as Figure 25)

3.5.7 Summary of key points

The analysis above shows that there is a very clear trend of demethylation increasing with increasing pre-treatment methylation level. This trend is apparent for both absolute and proportionate changes, although it is most clear for absolute changes. Once the trend is taken into account, the other three factors have little further impact on the level of demethylation, although there is some evidence that CpG density may have a small, additional effect.

3.6 Predictors of methylation change

Section 3.3 above identified that both cell lines and three of the patient samples experienced demethylation following treatment with DAC, whereas the other five patient samples experienced very little demethylation. For the former group, this section analyses the factors which may predict how the level of demethylation varies across different CpG sites.

3.6.1 Pre-treatment methylation level is the main predictor of demethylation

The results in section 3.5 suggest that the principal determinant of the extent to which a particular CpG site will be demethylated following treatment with DAC is the site's pre-treatment methylation level.

In order to formally confirm the conclusion reached in section 3.5, a linear regression analysis was performed. For this purpose, absolute beta value change was the predicted outcome, and pre-treatment beta value, gene location, CpG island region and CpG density (as measured by CpG count) were the candidate predictor variables. For the five members of group 1, the correlations (R^2) identified by this analysis are set out in table 5 below.

Variables used	HL60	KGIA	RILNIA	ROSBOS	ZPEA
Pre-treatment beta value alone	0.846	0.825	0.657	0.669	0.540
Pre-treatment beta value and CpG density	0.846	0.826	0.660	0.672	0.561
Pre-treatment beta value and gene location	0.846	0.826	0.658	0.671	0.543
Pre-treatment beta value and island region	0.846	0.826	0.658	0.671	0.551

Table 5 - R^2 values for associations of factors with absolute beta value change

For all five, this analysis shows that a significant proportion of the variation in demethylation levels between CpG sites is explained by pre-treatment methylation level, and that the other three candidate variables add very little to this association. The results are most striking for the two cell lines, which experienced the greatest levels of demethylation.

For the other five patient samples, the associations between each of the candidate variables and beta value change are very small (maximum R^2 of 12.8%).

If proportionate changes are considered instead of absolute changes, the levels of association, as set out in table 6 below, are lower but follow a similar pattern to table 5.

Variables used	HL60	KGIA	RILNIA	ROSBOS	ZPEA
Pre-treatment beta value alone	0.581	0.527	0.281	0.370	0.261
Pre-treatment beta value and CpG density	0.581	0.528	0.283	0.370	0.274
Pre-treatment beta value and gene location	0.581	0.528	0.281	0.370	0.263
Pre-treatment beta value and island region	0.582	0.528	0.282	0.370	0.269

Table 6 - R^2 values for associations of factors with proportionate beta value change

Given that absolute changes are correlated with pre-treatment methylation levels, one might expect that the correlation between proportionate changes and pre-treatment methylation levels would be lower. However, there is still a correlation meaning that proportionate changes are not independent of pre-treatment methylation levels. Also, the analysis using proportionate changes is possibly subject to greater distortion caused by measurement inaccuracy.

3.6.2 Logistic regression analysis of factors affecting demethylation

In order to further analyse the associations between the four candidate variables and level of demethylation, a logistic regression analysis was performed. For this purpose, the outcome to be predicted for each CpG site was whether the site was demethylated following treatment with DAC. Two separate thresholds for the absolute reduction in beta value were used to determine whether a site had been demethylated:

- **20%** - in the literature, a 20% cut-off has typically been used to test whether a site is demethylated - i.e. a site's beta value must have reduced by at least 20%. This is based on an analysis by Bibikova, which stated that a 20% absolute change could be detected with 99% confidence by the Illumina array.
- **10%** - as mentioned in section 2.5, this study has found that 99.5% of all pairs of replicate beta value measurements using the Illumina array were within 10% of each

other. Therefore, it could reasonably be argued that a 10% absolute change could be used as an alternative threshold.

As illustrated in section 3.3, using a 20% cut-off is very restrictive for all the patient samples (as only a very small proportion of CpG sites experienced demethylation of at least 20%), and the logistic regression analysis produced very small associations between all of the four candidate variables and demethylation. However, a 20% threshold is a lot less restrictive for the two cell lines (each having around 59% of sites which were demethylated by at least 20%), and for these samples the logistic regression found strong associations between demethylation and pre-treatment beta value (Nagelkerke's R^2 values of 0.748 and 0.729 for the cell lines HL60 and KGIA respectively). Introducing the other other three variables into the regression analysis only increased these values by very small amounts (maximum 0.001).

A 10% threshold is less restrictive, especially for patient samples RILNIA and ROSBOS. Table 7 below shows the associations (measured by Nagelkerke's R^2) for both cell lines and RILNIA, ROSBOS and ZPEA.

Variables used	HL60	KGIA	RILNIA	ROSBOS	ZPEA
Pre-treatment beta value alone	0.858	0.877	0.421	0.356	0.140
Pre-treatment beta value and CpG density	0.863	0.880	0.423	0.360	0.157
Pre-treatment beta value and gene location	0.858	0.877	0.423	0.359	0.143
Pre-treatment beta value and island region	0.861	0.879	0.422	0.358	0.148

Table 7 - Nagelkerke's R^2 values for associations of factors with beta value change

Whilst the associations for RILNIA, ROSBOS and, in particular, ZPEA are a lot smaller than those for the two cell lines, there is still a clear pattern of pre-treatment methylation level having the strongest association with demethylation, with the other three variables adding very little.

3.6.3 Stratification of average methylation change by pre-treatment methylation level

Table 8 below shows how the average absolute reduction in beta value, for each cell line and patient sample, varies according to pre-treatment beta value (grouped into 10% ranges).

	HL60	KGIA	RILNIA	ROSBOS	ZPEA	APAL	GALWIL	NALE	DONCO	CBUN
90%-100%	39.2%	36.0%	16.4%	15.4%	12.1%	-0.1%	0.3%	3.3%	1.2%	1.3%
80%-90%	32.0%	29.5%	12.0%	11.2%	7.5%	-0.7%	-0.8%	1.7%	0.4%	0.4%
70%-80%	27.1%	25.0%	9.7%	9.1%	5.4%	-1.7%	-2.0%	0.5%	-0.6%	-0.8%
60%-70%	21.0%	19.0%	6.9%	6.4%	3.0%	-2.3%	-2.4%	-0.4%	-0.8%	-1.4%
50%-60%	15.4%	14.5%	4.4%	3.5%	1.0%	-1.4%	-2.0%	-0.5%	-1.0%	-1.8%
40%-50%	10.7%	10.9%	2.7%	1.8%	0.1%	-0.6%	-2.0%	-0.8%	-1.2%	-2.2%
30%-40%	7.3%	7.6%	2.0%	1.0%	-0.8%	0.2%	-1.6%	-0.8%	-1.1%	-2.0%
20%-30%	3.8%	3.2%	0.7%	0.2%	-1.4%	0.7%	-1.0%	-0.7%	-0.7%	-1.2%
10%-20%	-0.7%	-0.7%	-0.6%	-0.6%	-1.2%	0.5%	-0.6%	-0.6%	-0.5%	-0.4%
0%-10%	-0.8%	-0.4%	-0.5%	-0.5%	-0.5%	0.1%	-0.3%	-0.5%	-0.5%	-0.3%

Table 8 - average reductions in beta values broken down by pre-treatment beta value groups

For all five members of group 1 this table shows a clear association between pre-treatment methylation level and the extent of demethylation following treatment with DAC.

3.6.4 Possible secondary association with CpG density

Section 3.5.6 suggested that there may be a secondary association of methylation change with CpG density for sites whose pre-treatment methylation levels are in the 40% to 60% range. In order to test this, a linear regression was performed for sites in this range. Taking the HL60 cell line as an example, the regression showed that for sites in this range, pre-treatment methylation level and CpG density had very similar correlations with the level of methylation change (R^2 values of 17.2% and 16.0% respectively). Similar results were obtained for the KGIA cell line and patient samples RILNIA, ROSBOS and ZPEA. The analysis also showed that there was no interaction between CpG density and pre-treatment methylation level.

3.6.5 Summary of key points

The key result from this analysis is that, of the factors analysed in this study, pre-treatment methylation level is by far the most important in predicting the extent to which a particular CpG site will be demethylated following treatment with DAC. This is observed for both absolute and proportionate changes, although the correlations are lower for proportionate changes (possibly in part because of the potential inaccuracy in using proportionate changes). There also appears to be a secondary association with CpG density for CpG sites whose pre-treatment methylation levels are in the range 40% to 60%. However, for the patient samples in particular, a lot of variation in demethylation levels is still unexplained.

4. Discussion

The analysis showed considerable variation in the level of demethylation between samples. In particular, the two cell lines and three of the patient samples were demethylated to a much greater extent than the other five patient samples. Within samples, CpG demethylation was found to vary according to gene location, CpG density, CpG island region and pre-treatment methylation level. Multivariate regression analysis showed that of the factors investigated, the principal determinant of demethylation at an individual CpG site was the pre-treatment methylation level (high pre-treatment methylation implies high reduction and vice-versa).

The principal correlation between pre-treatment methylation level and level of demethylation following treatment was observed when both absolute and proportionate methylation changes were considered. Although the correlation was lower for the latter case, it was still evident meaning that proportionate changes in methylation are not independent of pre-treatment methylation level. It should also be noted that the use of proportionate changes is potentially subject to distortion caused by measurement inaccuracy.

The level of demethylation also appeared to be inversely correlated to CpG density. However the stratification by CpG density and starting beta value clearly demonstrated that the latter has by far the greater impact. CpG density still has an effect, but this is much smaller and tends to be concentrated in the 40% to 60% pre-treatment beta value range.

Once the effect of pre-treatment methylation level is allowed for, there is no obvious association of beta value change to gene location, whilst, as might be expected, there is a high correlation of CpG density to CpG island region, meaning that the latter does not appear to have any additional, independent influence.

It is possible that the primary association with starting beta value can be at least partially explained by the effect known as "regression to the mean" - i.e. some of the higher pre-treatment beta values may have been overstated (by measurement error), so a subsequent re-measurement is likely to be lower (i.e. nearer the correct value) and vice-versa[18]. Cost considerations meant that it was not possible to process technical replicates of the AML samples using the Illumina array, so the precision of the measurements could not be tested directly. However a further analysis was performed using the same array and in the same laboratory on a completely independent data set which did include technical replicates. Comparison of the methylation measurements between replicates suggested that regression to the mean is not an issue. Also, the distribution of results for the other five patient samples gave no indication that regression to the mean is having a material impact on the results.

The key, most directly comparable previous research into this area is covered in the paper by Hagemann et al[11]. That study also found that demethylation following treatment with DAC is dependent on pre-treatment methylation level, and also that CpG sites in CpG islands were less likely to be demethylated. However, this project was able to provide a more refined analysis - for example, the Hagemann paper was based on the older version of the Infinium technology[19] which, in particular, interrogated around 28,000 CpG sites compared with over 480,000 sites analysed by the newer technology used for this project.

It is also important to stress that the analysis in the Hagemann paper was based exclusively on cell lines. It may be unwise to extrapolate results observed from cell lines to human samples. Indeed, this project has shown that demethylation in the two cell lines was greater than in all eight patient samples, and that there was substantial variation across the patient samples.

The results are also consistent with other research by Rubinstein et al[20], which investigated the effect of DAC on gene expression levels in melanoma cell lines and looked at the associations with pre-treatment methylation levels and CpG content. Rubinstein found that gene promoters with high pre-treatment methylation levels and intermediate CpG content appeared most susceptible to up-regulation following treatment with DAC.

Recent research by Yan et al[21], which used methylated DNA capture in combination with next generation sequencing to investigate the effect of DAC treatment on genome-wide methylation levels of 16 AML patients aged over 60, also found that the extent of demethylation appears to be correlated to pre-treatment methylation levels. Consistent with this project, and confirming the heterogeneity which exists between patients, the Yan study also showed considerable variation in demethylation between samples. There was significant demethylation observed in the nine patients who achieved complete remission following treatment with DAC, whereas the other seven patients who did not achieve complete remission showed no significant changes in methylation.

However, the Yan study also showed that demethylation occurred mainly in CpG island regions, which is in contradiction to the results of this project. Hence one area of future work would be to investigate the reasons for this apparent discrepancy.

Another area where further work should be pursued is to look at the mechanisms by which DAC causes demethylation. The Hagemann paper[11] includes some results on this. Firstly, it identified that CpG sites associated with genes involved in the Polycomb Repressive Complex 2 appeared to be more resistant to demethylation than other CpG sites. Secondly, it identified that demethylation-sensitive and demethylation-resistant CpG sites showed complementary enrichment of certain transcription factor binding sites. Recent research by Metzeler et al[22]

has also linked reduced expression of DNMT3A in AML patients to response to DAC. Hence these are areas which would be worth investigating further.

There are a number of areas in which this project could be improved. Firstly, one potential cause of the variation in demethylation across the samples is the timescale used. The post-treatment methylation measurements were performed five days after treatment for all samples. Hence the results simply reflect a snapshot of demethylation levels at a single point in time. Primary samples grown *in vitro* do not all proliferate at the same rate. The variation in demethylation among the samples is likely to reflect these differences in proliferation, with the faster dividing cells demethylating to a greater extent than the slower growing cells. It is possible that if further measurements had been taken at different time-points, then different methylation patterns would have been observed.

Secondly, because AML is rare in children, it is difficult to obtain samples from paediatric patients. Hence the analysis was based on a small number of samples, which means that any conclusions are potentially biased by the particular characteristics of the samples and would need to be confirmed in a larger group of samples.

Finally, the Infinium HumanMethylation450 BeadChip technology is relatively new, having been released in 2011. There are a number of papers which attempt to investigate the accuracy of the technology[23,24,25,26], but there is still no clear consensus on the best approach for downstream analysis of the output. For example, Touleimat[26] considered eight different methods for processing the raw Illumina data, and compared the results for a small sample of CpG sites against a separate analysis using pyrosequencing, but did not show conclusively that any one method was superior to the others. Hence, whilst there is no obvious evidence of inaccuracy, and this was corroborated by the pyrosequencing analysis used for this project, it

would be desirable for further investigation of the technical accuracy of the Illumina array to be carried out.

In conclusion, this project has shown that pre-treatment methylation level appears to be the main determinant of the extent to which an individual CpG site will be demethylated following treatment with DAC. However, the analysis also showed that the determinants identified were in themselves insufficient to explain all of the variation in demethylation observed across study samples. Further work with larger sample sizes is required to validate this result and to help elucidate the mechanisms by which DAC operates.

List of references

1. Mrozek K, Heerema NA and Bloomfield CD (2004) Cytogenetics in acute leukemia. Blood Rev 18, 115-136

2. Mrozek K, Marcucci G, Paschka P and Bloomfield CD (2007) Clinical relevance of mutations and gene-expression changes in adult acute myeloid leukemia with normal cytogenetics: are we ready for a prognostically prioritized molecular classification? Blood 109, 431-448

3. Jiang Y, Dunbar A, Gondek LP, Mohan S, Rataul M et al (2009) Aberrant DNA methylation is a dominant mechanism in MDS progression to AML. Blood 113, 1315-1325

4. Jones PA (1999) The DNA methylation paradox. Trends in Genetics 15(1), 34-37

5. Lyko F and Brown R (2005) DNA Methyltransferase inhibitors and the development of epigenetic cancer therapies. Journal of the National Cancer Institute, 97(20), 1498-1506

6. Jones PA and Baylin SN (2007) The epigenomics of cancer. Cell 128(4), 683-692

7. Lugthart S, Figueroa M, Bindels E, Skrabanek L, Valk PJM et al (2011) Aberrant DNA hypermethylation signature in acute myeloid leukemia driected by EVI1. Blood 117, 234-241

8. Buchi F, Spinelli E, Masala E, Gozzini A, Sanna A et al (2012). Proteomic analysis identifies differentially expressed proteins in AML1/ETO acute myeloid cells treated with DNMT inhibitors azacytidine and decitabine. Leukemia Research 36, 607-618

9. Christman JK (2002). 5-Azacytidine and 5-aza-2-deoxycytidine as inhibitors of DNA methylation: mechanistic studies and their implications for cancer therapy. Oncogene; 21:5483–5495

10. Laird PW (2010) Principles and challenges of genome-wide DNA methylation analysis. Nature Reviews 11, 191-203

11. Hagemann S, Heil O, Lyko F and Brueckner B (2011) Azacytidine and Decitabine Induce Gene-Specific and Non-Random DNA Demethylation in Human Cancer Cell Lines. PLoS One, 6-3, e17388

12. D Takai and PA Jones (2002) Comprehensive analysis of CpG islands in human chromosomes 21 and 22. Procedures of the National Academy of Science USA 99, 3740-3745

13. Bibikova M, Barnes B, Tsan C, Ho V, Klotzle B et al (2011) High Density DNA methylation array with single CpG site resolution. Genomics 98, 288-295

14. Irizarry RA, Ladd-Acosta C, Wen B, Wu Z, Montano C et al (2009) The human colon cancer methylome shows similar hypo- and hypermethylation at conserved tissue-specific CpG island shores. Nature Genetics 41, 178-185

15. Wang D, Yan L, Hu Q, Sucheston LE, Higgins MJ et al (2012) IMA: an R package for high-throughput analysis of Illumina's 450K Infinium Methylation data. Bioinformatics 28, 729-730

16. Nagelkerke NJD (1991) Miscellanea - A note on the general definition of thee coefficient of determination. Biometrika 78:3, 691-692

17. Leonard S, Wei W, Anderton J, Vockerodt M, Rowe M et al (2011) Epigenetic and transcriptional changes which follow Epstein-Barr Virus infection of germinal center B cells and their relevance to the pathogenesis of Hodgkin's Lymphoma. Journal of Virology 85(18), 9568-9577

18. Barnett AG, van der Pols JC and Dobson, AJ (2005) Regression to the mean: what it is and how to deal with it. International Journal of Epidemiology 34, 215-220

19. Bibikova M, Le J, Barnes B, Saedinia-Melnyk S, Zhou L et al (2009) Genome-wide DNA methylation profiling using Infinium® assay. Epigenomics 1, 177-200

20. Rubinstein JC, Tran M, Ma S, Halaban R and Krauthammer M (2010) Genome-wide methylation and expression profiling identifies promoter characteristics affecting demthylation-induced gene up-regulation in melanoma. BMC Medical Genomics 3:4, 1-9

21. Yan P, Frankhouser D, Murphy M, Tan H-H, Rodriquez B et al (2012) Genome-wide methylation profiling in decitabine-treated patients with acute myeloid leukemia. Blood pre-published online July 11, 2012; doi:10.1182/blood-2012-05-429175

22. Metzeler KH, Walker A, Geyer S, Garzon R, Kilsovic RB et al (2012) DNMT3A mutations and response to the hypomethylating agent decitabine in acute myeloid leukemia. Leukemia 26, 1106-1107

23. Sun Z, Chai HS, Wu Y, White WM, Donkena KV et al (2011) Batch effect correction for genome-wide methylation data with Illumina Infinium platform. BMC Medical Genomics, 4:84, 1-12

24. Dedeurwaerder S, Defrance M, Calonne E, Denis H, Sotiriou C and Fuks F (2011) Evaluation of the Infinium Methylation 450K technology. Epigenomics 3(6), 771-784

25. Roessler J, Ammerpohl E, Gutwein J, Hasemeier B, Anwar SL et al (2012) Quantitative cross-validation and content anaysis of the 450K DNA methylation array from Illumina, Inc. BMC Research Notes, 5:210

26. Touleimat N and Tost J (2012) Complete pipeline for Infinium Human Methylation 450K BeadChip data processing using subset quantile normalization for accurate DNA methylation estimation. Epigenomics 4(3), 325-341

Evaluating the predictive ability of "omic"-based biomarker models in human cancer

by

Robert John Hollows

This project is submitted in partial fulfilment of the requirements for the award of the MRes in Biomedical Research

School of Cancer Sciences

College of Medical and Dental Sciences

University of Birmingham

September 2012

Abstract

In recent years new "omic"-based technologies, such as microarrays, have been used to create many novel biomarker models for predicting outcomes in human cancer. However, only a fraction of these models have been put into actual clinical use. For models to have proven value, they must be shown to be generalisable to the wider population away from the data sets used to create them. To achieve this models must be properly validated. Various studies have considered how such validation should be performed.

In this study, a survey has been undertaken of 100 recent papers which have claimed to have validated "omic"-based biomarker models. The purpose of the study was to compare actual validation methodologies being used against best practice as set out in the literature. The results show that there are considerable deficiencies in the way that validation is undertaken, in particular with regard to sample sizes which are too small, inappropriate handling of data and over-reliance on validation methods which do not use genuinely independent data. Also, there is a disappointing shortage of studies undertaking independent validation of models constructed by other research teams.

In conclusion, more emphasis is required on the proper validation of biomarker models.

Acknowledgements

I would like to thank Professor Jon Deeks and Dr Richard Riley for their valuable assistance and guidance with this project.

Table of contents

List of figures

List of tables

1. Introduction

This project surveyed recent scientific literature concerning the validation of models created for predicting outcomes in human cancer using biomarker data derived from the various "omic"-based technologies (i.e. genomics, transcriptomics, proteomics, metabolomics and epigenomics).

1.1 Background

In the last few years, technological advancements (such as microarrays) have made it possible for researchers to quickly and efficiently analyse large amounts of biological data. One of the key uses these technologies have been put to is to try to make diagnostic or prognostic predictions in relation to outcomes in cancer (e.g. probability of getting cancer, probability of metastasis etc) based on "omic" biomarker data extracted from actual human samples.

Many scientific papers have been published claiming to have produced novel "omic"-based models for use in cancer prediction[1]. However, only a very small proportion of these models have actually been approved for clinical use[2]. Not only is this an inefficient use of scientific resources, but it could also be argued to be detrimental to overall scientific knowledge as the literature becomes cluttered with large numbers of papers which have minimal, if any, practical usefulness.

1.2 The key problems

There are several potential issues which can limit the usefulness of a newly-created biomarker model. These issues can be broken down into the following broad categories[2,3]:

1. Incorrect or inappropriate use of technology by which the raw data are extracted
2. Inappropriate experimental design

3. Inappropriate mathematical / statistical techniques used for analysing the raw data and creating the model
4. Lack of suitable validation of the model once it has been created
5. Lack of translatability into actual clinical use

This project concentrates on the fourth of these areas, whilst also touching on the third and fifth.

The key issue which this project considers is that, whilst a model might appear to have very good predictive value when applied to the data on which it is based, it is essential that the results are confirmed in a separate analysis using an independent data set. A check on the results provides a test of whether the model is generalisable to the wider population, or whether it is overfitted to the data on which it is based - i.e. rather than identifying genuine biological differences in the selected biomarkers which are present in the whole population, it also simply reflects particular characteristics of the samples used to create it[1]. Overfitting is a common problem with predictive models, and although it can be improved by using large sample sizes[4], often cost issues and the scarcity of suitable data make this difficult to achieve in practice.

1.3 The solution - validation

In order to overcome the problem of overfitting, it is imperative for a model to be validated. Validation should ideally be performed on (at least one, but preferably several) completely independent, external data sets - hereafter referred to as "external validation"[3]. Such validation will test the generalisability of the model to the wider population.

In practice, other validation techniques are also used which do not make use of external data sets but are instead based on the data set used to create the model - hereafter referred to as "internal validation"[3]. This can involve the splitting of the data into a "training" set used to create the model and a "validation" set used to test its performance. The crucial point here is that both sets of data have come from the same source, and are therefore not genuinely independent, even if they contain different samples.

Internal validation also includes such techniques as cross-validation and permutation testing. Cross-validation is a frequently used method for assessing the predictive accuracy of a model without using any additional data. Typically, the entire training dataset is split randomly into two subsets. The first, usually larger, of these subsets is used to train the model using exactly the same methodology as was used for the entire data set. This (sub-) model is then tested on the second subset of data. The process is repeated several times and the average performance across all random splits of the data is used to assess the overall predictive performance of the model.

The relative sizes of the subsets can vary - for example, at one extreme (known as leave-one-out-cross-validation, or "LOOCV") the training subset includes all bar one of the members of the full data set, and is used to train a model which is then tested on the one excluded member, this process then being repeated with every member of the full data set treated in turn as the test member. Other variants used are known as k-fold where the full data set is randomly split into k subsets of equal size, and then each of these subsets in turn is used as the testing subset with the other k-1 used together as the training subset.

Another technique often used to assess the statistical significance of a model's performance (i.e whether a model is predictive, rather than how accurate its predictions are), is the use of

permutation testing. Once a model has been used and the results obtained, the significance of the results can be estimated by repeatedly, randomly reassigning all the samples to the observed outcomes and recalculating the relevant performance statistics on the permuted data. If this process is performed a large number of times, then the significance of the model performance can be estimated by ranking all the permuted results and finding where the actual result based on the actual, unpermuted data, falls within these rankings. Based on 10,000 permutations for example, if the actual result falls within the extreme 500 most results then it could be said that the result is significant at the 5% level.

Previous research has found that internal validation techniques, whilst preferable to undertaking no validation at all, are not as successful as external validation at accurately gauging the generalisability of a model - the key issue being that they tend to overestimate the performance of the model[5].

1.4 How to assess model performance

Whatever validation methodology is used, there are two critical areas of performance of the model which should be tested. These are "calibration" and "discrimination"[3]. Calibration compares the frequency of predicted and actual outcomes, whilst discrimination measures the model's ability to distinguish between patients with and without a particular outcome.

There are various statistical techniques available to assess both of these measures. For example, calibration can be assessed by comparing observed proportions of events against predicted probabilities. For discrimination, the standard statistical measures of sensitivity and specificity can be calculated, along with the area-under-the-curve (AUC) of the receiver operating characteristics curve (ROC).

For a model to actually be approved for clinical use, a cost-benefit analysis of applying the model in practice needs to be made[2]. Such an analysis would compare the benefits from making a correct prediction using the model with the costs of making an incorrect decision. For example, diagnosing the presence of a particular cancer may result in some form of exploratory surgery. Clearly if the diagnosis is correct, then the surgery is worthwhile and there is a huge benefit to the patient. However, there will be a (hopefully very small) proportion of patients for whom the diagnosis is incorrect, leading to unnecessary surgery. If the overall costs of unnecessary surgery outweigh the benefits of correct diagnoses, then it might be considered inappropriate to use the model even if a correct diagnosis is of huge benefit to the individuals concerned.

1.5 Aims of project

Given the importance of proper validation of newly created biomarker models, the key aims of this project were to:

- systematically review recent scientific literature to survey the validation methodologies used in practice. The survey was focused on models which were created using "omics"-based data, which were intended to be used for predicting outcomes relating to human cancer, and which had been subject to validation in some form;
- document and summarise the ways in which validation was undertaken - e.g. methodology used, sample sizes;
- compare actual practice with what is known about validation methodology best practice from relevant scientific literature;

- where the information was available, compare measures of model performance based on different validation methodologies.

A literature review was undertaken to capture the most recent 100 scientific papers which involved the validation of such biomarker models. Relevant items of data were extracted from each of these papers, and then analysed.

2. Data and methods

There were four key stages to this project:

- literature search to identify potential papers
- filtering process to select most recent 100 papers meeting inclusion criteria
- establishment of database of answers to pre-specified questions
- analysis of database and comparison of findings against best practice

Each of these stages is considered in turn below.

2.1 Literature search

The first part of this project was the literature search. The purpose of the search was to find papers which were relevant to the project according to the following three criteria:

1. They each involved the validation of a predictive, mathematical model.
2. The primary outcome being predicted involved some aspect of human cancer.
3. The model used "omic"-based biomarkers.

The search was performed on 7 May 2012 using the biomedical literature database Medline, which was accessed on-line via the University of Birmingham's e-Library facility. Table 1 below shows how each of the three criteria mentioned above were allowed for in the search command (note: at this stage, validation was not included in the search as this was found to be too restrictive):

Criterion	Search string	Location searched	Medline search command
1.	"predict*" or "prognos*"	Title	"predict*".m_titl. OR "prognos*".m_titl.
2.	"cancer" OR "cancer" or "Neoplasms" (all sub-headings)	Title Keywords	cancer.m_titl. OR cancer.mp.or exp Neoplasms/
3.	"genom*", "epigenom*", "transcriptom*", "proteom*", "metabolom*" or "gene*"	Title	"genom*".m_titl. OR "epigenom*".m_titl. OR "transcripto*".m_titl. OR "proteom*".m_titl. OR "metabolom*".m_titl. OR "gene*".m_titl.
The final search command combined all of these three commands with the "AND" operator.			

Table 1 - details of literature search command using Medline biomedical database

2.2 Filtering process

The search using the above commands identified 3,190 papers. A filtering process was then undertaken, applied to the papers in reverse chronological order, until 100 papers which satisfied all requirements had been selected. The requirements for each paper were:

- It was written in English.
- It was not a review article.
- It had an abstract.
- The abstract included clear indication that all of the three criteria mentioned in section 2.1 were met - in particular that validation had been performed.
- The paper was available for download via the University of Birmingham's e-Library.

2.3 Establishment of database

Once 100 papers had been identified, the next stage was to extract relevant information from each of the papers and to store this in a database. A set of questions was devised, as outlined below, and then each of the papers was studied in turn, with the answers to each question being recorded in a database created using Microsoft Excel.

The questions applied to each paper were subdivided into four main categories, which were:

- Background information about the paper.
- Information about the model-building process.
- Information about any internal validation process which had been undertaken.
- Information about any external validation process which had been undertaken.

The questions which were asked under each category were as follows:-

2.3.1 Background information

A small number of questions were asked to gain an understanding of the background of the 100 papers. These are set out in table 2 below:

1.	In which year was the paper published?
2.	In which country was the primary author based?
3.	What type of cancer was the subject of the paper?
4.	The paper was based on which of the "omics"?
5.	Did the paper cover both the creation and validation of a model, or just the validation of an existing model?
6.	Was the model to be used for diagnostic, prognostic or predictive purposes?

Table 2 - questions relating to background information

The final question was used to identify whether the purpose of the model was to:

- make a diagnosis - e.g. whether someone had cancer; or
- make a prognosis - e.g. how long would someone already with cancer be expected to live; or
- make a prediction of the impact of treating someone in a particular way - e.g. how long would someone be expected to live if they were treated using chemotherapy?

The possibility that some models might include more than one of these options was allowed for. In this case, the questions in the other three categories, as described below, were applied separately for each aspect of the model.

2.3.2 Model-building process

The next part of the questionnaire concerned aspects of the model-building process. These questions were used to identify key issues which have been shown to affect the usefulness of biomarker-based models and to enable comparison of any performance statistics with validation results. The questions considered how the data were gathered, sample sizes, the numbers and types of biomarkers, details of the final model, and key performance measures.

The questions posed are set out in table 3 below:

1.	Was the model based on a prospective or retrospective study? A model was deemed to be based on a retrospective study if it was clear from the paper that it was based on the use of an old data set. Alternatively, it was deemed to be prospective if it was clear that patients had been genuinely prospectively recruited and the data had only recently been extracted. If there was any doubt an answer of "unclear" was recorded.
2.	If prospective, was the study planned for use in the model-building exercise, or did the model-building make opportunistic use of a study originally designed for another purpose? If there was any doubt then an answer of "unclear" was recorded.

3.	What was the key outcome considered by the model (e.g. death from any cause, recurrence-free survival)?
4.	What type of biomarker variable was used in the model - e.g. was it a continuous, numerical value (such as gene expression level), or a binary amount (such as presence of absence of a particular genetic mutation)?
5.	In the initial selection of biomarkers to include in the model, was any explicit adjustment made to control false discovery rates?
6.	Was a completely separate data set used solely for the purposes of conducting an initial filtering of biomarkers to be used in the subsequent model-building?
7.	How many biomarkers were used in the model-building exercise?
8.	How many biomarkers were used in the final model?
9.	Samples from how many patients were used in the model-building exercise?
10.	Were any patient samples removed from the initial data set because of unavailable information?
11.	How many "outcomes" were observed amongst the patients (e.g. if the key outcome under consideration was death from any cause, how many patients died?)
12.	Was the validation data set (if there was one) used in any part of the model-building exercise?
13.	Was the final rule created by the model explicitly stated?
14.	Did the model split patients into groups (e.g. high or low risk), and, if so, how many?
15.	What method was used to determine the split into groups?
16.	Was the calibration of the model measured?
17.	If so, what method was used?
18.	Were specificity and sensitivity measured?
19.	If so, what were their values?
20.	Was the AUC measured?
21.	If so, what was its value?
22.	Were any other discrimination measures reported?
23.	Was any decision analysis performed?

Table 3 - questions relating to model-building process

2.3.3 Internal validation process

The next section of the questionnaire included the questions in table 4 below which were asked to identify key issues concerning any internal validation performed on each model:

1.	Was internal validation with a split of the data set (into training and validation parts) used?
2.	If so, how was the data set split?
3.	How many patients were in the validation part of the data set?
4.	And how many outcomes were observed amongst these patients?
5.	Was any cross-validation used?
6.	If so, what technique was used (e.g. leave-one-out-cross validation, two-fold etc)
7.	And what was the key purpose of the cross-validation?
8.	Did the cross-validation include the biomarker selection process?
9.	Was permutation testing used?
10.	If so, how many permutations?
11.	And what was the key purpose of the permutation testing?
12.	Was the calibration of the model measured?
13.	If so, what method was used?
14.	Were specificity and sensitivity measured?
15.	If so, what were their values?
16.	Was the AUC measured?
17.	If so, what was its value
18.	Were any other discrimination measures reported?

Table 4 - questions relating to internal validation process

The key purposes of asking these questions were to identify how commonly each type of internal validation was used, how they were applied in practice, the sample sizes on which the validation was undertaken (where appropriate), and the frequency of reporting of key performance measures and their actual values for comparison with corresponding figures from the model-building process.

2.3.4 External validation process

Finally, the following similar group of questions, as set out in table 5 below, were asked to identify key issues concerning any external validation performed on each model:

1.	Was any external validation performed?
2.	If so, was the validation data set based on a prospective or retrospective study?
3.	If prospective, was the study planned for use in the validation exercise, or did the validation make opportunistic use of a study originally designed for another purpose?
4.	How many independent data sets were used?
5.	Samples from how many patients were used?
6.	Were any patient samples removed from the initial data set(s) because of unavailable information?
7.	How many "outcomes" were observed amongst the patients?
8.	Were the patients taken from the same location as those used for the model-building?
9.	Was the validation data gathered by someone other than the research team who originally produced the model?
10.	Was the calibration of the model measured?
11.	If so, what method was used?
12.	Were specificity and sensitivity measured?
13.	If so, what were their values?
14.	Was the AUC measured?
15.	If so, what was its value?
16.	Were any other discrimination measures reported?

Table 5 - questions relating to external validation process

2.4 Analysis of database

Once the database had been constructed, the final stage of the project was to analyse the data, identify the principal findings and compare against recognised best practice. Key items included in the analysis were the proportions of studies using the different types of validation, investigation of sample sizes used and methods used for collecting and handling data. Where possible, an assessment was made of the effect of validation on reported model performance, and how this compared across alternative validation methodologies.

The results are set out in Section 3. All charts were produced using Microsoft Excel.

3. Results

The 100 papers selected for this project are listed in the Appendix. The flowchart in figure 1 below outlines the selection process:

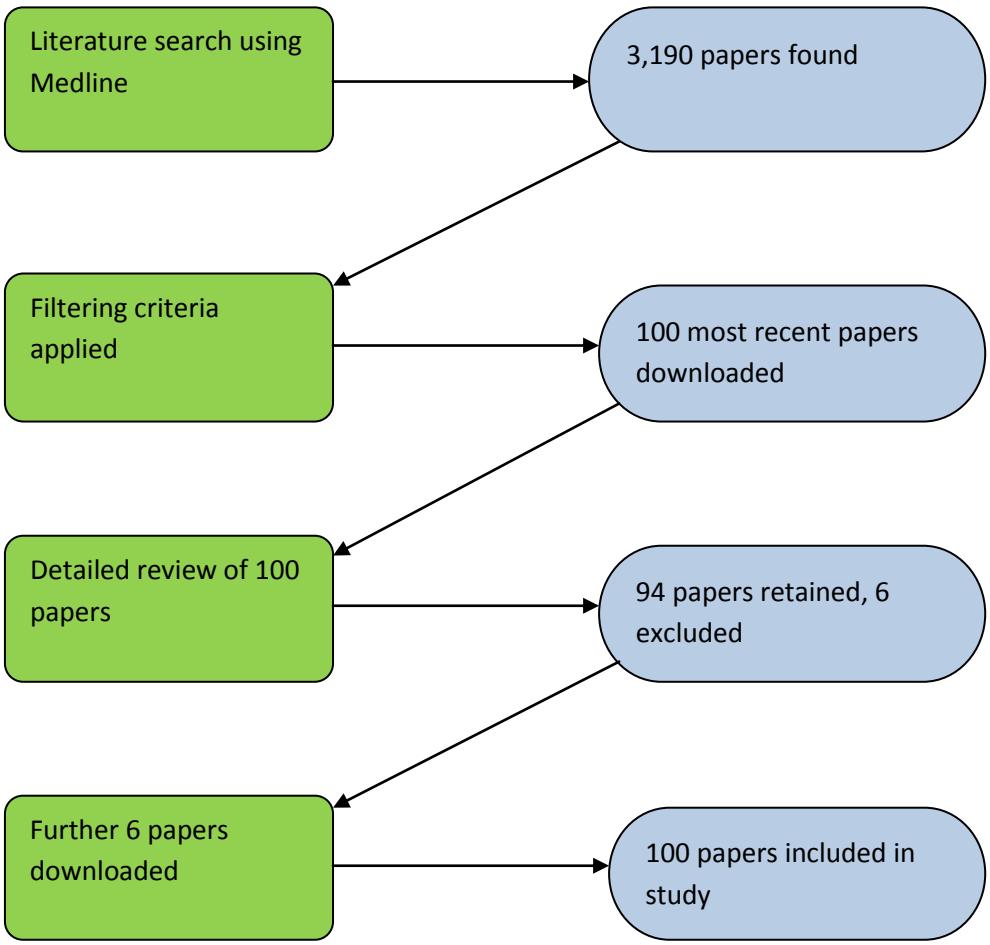


Figure 34 - flowchart of process for selecting 100 papers

Orientation of analysis

The analysis is set out in the following order.

Firstly, section 3.1 covers the questions concerned with background information about the 100 studies. Section 3.2 then covers aspects of the model building process, and sections 3.3 and 3.4 deal with internal and external validation processes respectively. Finally, section 3.5

analyses model performance measures and, where possible, compares these between those reported as part of the model-building exercise and those reported as part of the validation process.

3.1 Background information

The 100 papers surveyed covered two entire years (2010 and 2011) and two partial years (2009 and 2012). The numbers of papers from each year were five from 2012, 38 from 2011, 41 from 2010 and 16 from 2009.

A key observation is that the vast majority of papers (95) concerned both the construction of a biomarker model along with the subsequent validation of that model. Only five papers purely concerned validation of a previously constructed model (refs S44, S53, S76, S85, S96).

The full range of "omics" was covered, although transcriptomics was by far the largest category, being the sole subject of 78 of the papers. Genomics, proteomics, metabolomics and epigenomics were covered in 10, five, three and three papers respectively, whilst one paper (ref S47) considered the combination of both genomics and transcriptomics.

60 of the papers concerned purely prognostic models (as defined in section 2), whilst 23 were purely predictive and 10 were purely diagnostic. Four covered both diagnostic and prognostic models, two covered both prognostic and predictive models and one (ref S30) covered all three types. Hence, 108 models were included in total.

30 different cancer types were covered. The most common are shown in figure 2 below.

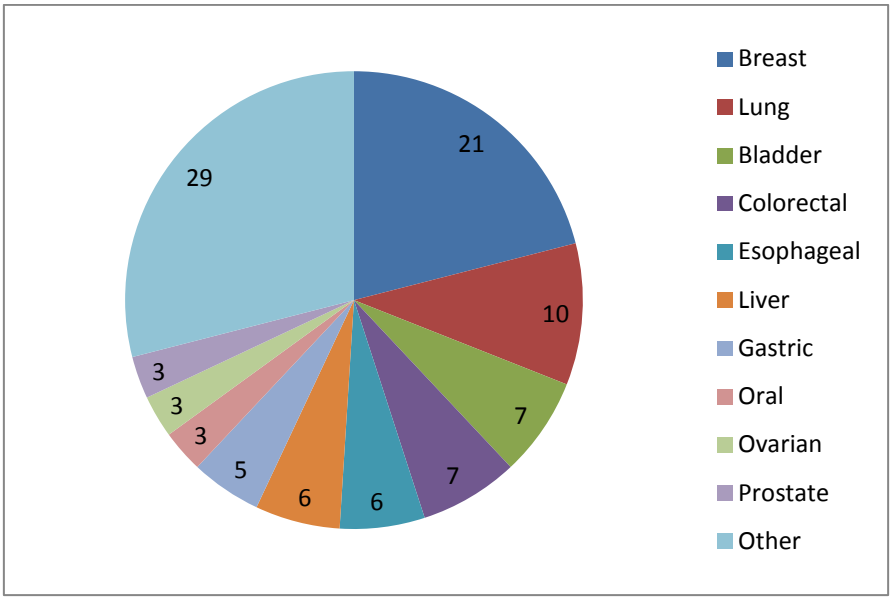


Figure 35 - types of cancer covered in the 100 papers

The primary authors of the papers came from 21 different countries. The USA was by far the most frequent, at 34, followed by Japan and South Korea with nine, and China with seven.

3.2 Model building

3.2.1 Type of data used

28 models were based on prospectively gathered data (as defined in section 2), 68 used retrospective data and for the remainder this determination was unclear. 12 of the prospective studies were planned, three were opportunistic and 13 could not be determined.

The vast majority (92%) of the models used continuous variables, whilst the remainder used binary variables with two outcomes.

3.2.2 Definition of outcome being investigated

Of the 15 diagnostic models covered, six were concerned with predicting whether an individual had cancer, and another three concerned existence of metastasis. The other six studies covered a variety of alternative outcomes.

The majority of the 67 prognostic models concerned either overall survival (28) or disease-free survival (29). Another six studies used metastasis-free survival as the primary outcome. Similarly, of the 26 predictive models, five were concerned with overall survival, seven with disease-free survival and three with metastasis-free survival. The primary outcome of a further nine was related in some way to the response to treatment (primarily chemotherapy).

3.2.3 Biomarker selection process

In the selection process for determining which biomarkers would be included in the model, only 15 studies used some form of adjustment to control false discovery rates on account of multiple testing. In 78 cases such an adjustment was clearly not made, whilst for the remainder the determination was unclear.

12 studies used a completely separate data set to perform an initial filtering of biomarkers to be included in subsequent model building. The range of sample sizes used, where this could be determined (10 out of 12 cases), was 3 to 297 with an average of 92. On the other hand, seven studies (refs S12, S17, S25, S26, S35, S50 & S92) actually used the validation data at some point in the model-building process (thereby destroying the independence of the validation data).

In 14 cases the final number of biomarkers used in the model was unclear. One extreme model (ref S88), which used transcriptomic data to make prognoses about survival of glioma patients, was based on a principal components analysis of 5,000 biomarkers. The distribution of the numbers of biomarkers used is shown in figure 3 below.

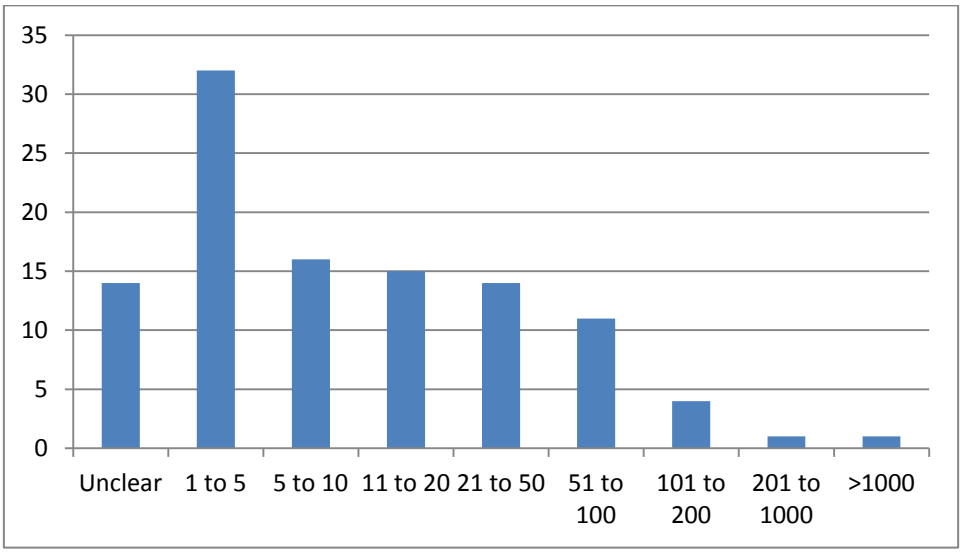


Figure 36 - distribution of numbers of biomarkers in final model

3.2.4 Patient / sample data

In 36 studies some patients / samples were removed from the full, starting data set as a result of missing information. This was also suspected for another three cases, although the evidence was unclear.

The number of patients used for the model-building process ranged from 13 to 1,125, with an average of 131. Just over one half of the models were constructed using 100 patients or less.

The number of outcomes (e.g. death) amongst the patients ranged from 2 to 187, with an average of 41, in those cases which provided this information. However, in 38 cases the number of outcomes could not be determined. When it could be calculated, the average ratio of number of outcomes to number of biomarkers considered was only 2.8, and in only four cases (refs S15, S46, S58, S86) did this ratio exceed 10.

3.2.5 Final model

The formula / rule underlying the final, created model was explicitly stated (or sufficient information to deduce it was provided) in only 34 studies. A good example of this is reference S12, which set out the precise formula for a model to predict the risk of myeloma patients developing bone disease when treated with bisphophonates.

92 studies used the model created to categorise patients into two risk groups (i.e. high and low risk). Seven models split patients into more than two groups (maximum six), whilst four models (refs S15, S32, S39, S47) used a continuous measurement with no splitting - for example, in reference S39 a model was created to predict the risk of an individual with oral premalignant lesion developing oral cancer . A variety of different methods was used to determine groups, and in 41 cases the methodology used was unclear.

3.2.6 Type of validation

The proportions of studies which used validation based on internal (i.e. same as model-building) data sources or genuinely external data are set out in figure 4 below:

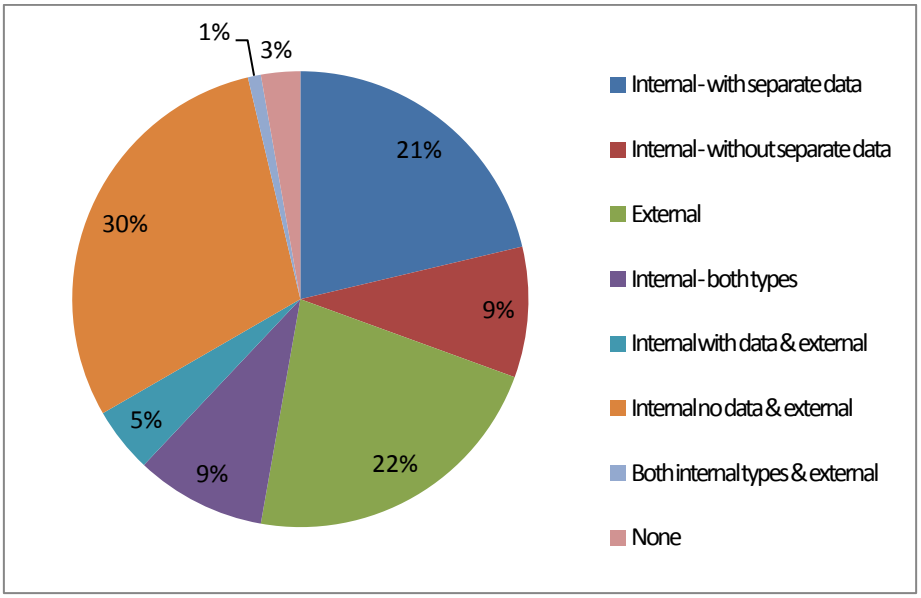


Figure 37 - breakdown of types of validation

In total, 62 models were validated using external validation, whilst internal validation was used for 81 models.

3.2.7 Summary of key points

The majority of models were based on retrospectively gathered data. Most used continuous variables but dichotomised patients into two risk groups. In only one third of studies was it possible to deduce the actual final rule applied by the model. There was a large range in the number of patients used and in the number of biomarkers contained in the final model. The ratios of outcomes to numbers of biomarkers were small. Internal validation was used more often than external validation.

3.3 Internal validation

Internal validation can be broken down into three different types:

1. validation using different patient / sample data from the same source
2. cross-validation with no separate data used
3. permutation testing with no separate data used

For those studies which used some form of internal validation, the breakdown between these three types is set out in figure 5 below.

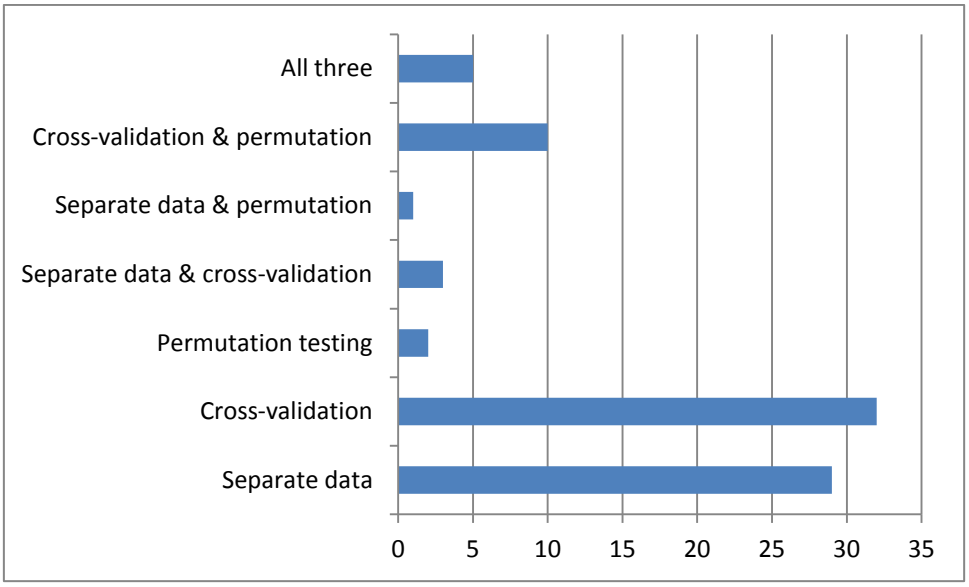


Figure 38 - distribution of different types of internal validation

3.3.1 Cross-validation techniques

The breakdown of the different ways in which cross-validation was utilised along with the purposes of the cross-validation are set out in figures 6 and 7 below.

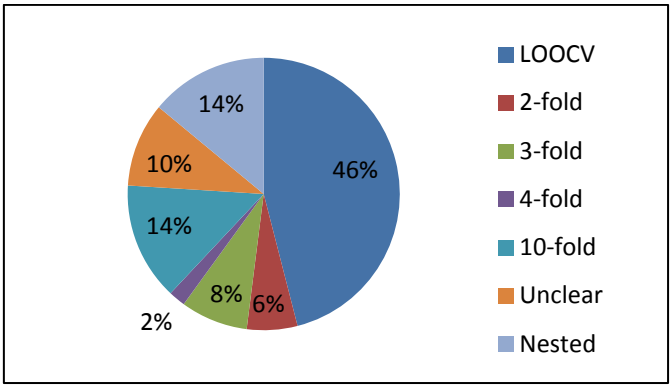


Figure 39 - distribution of different types of cross-validation

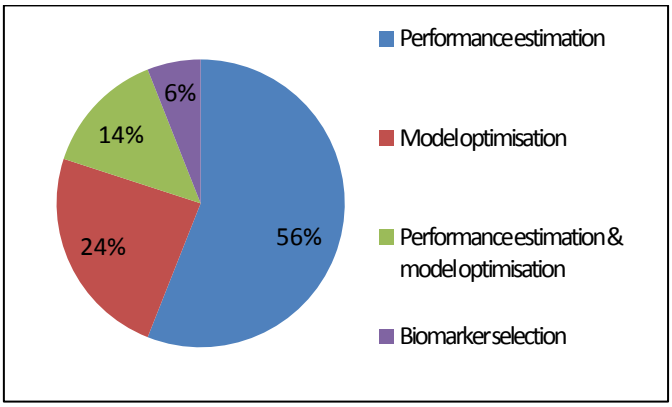


Figure 40 - distribution of different purposes of cross-validation

LOOCV was by far the most popular technique used, followed by a variety of different k-fold methods. Seven studies used a "nested" approach, whereby two separate cross-validation loops were used, one operating within the other.

The key reason for using cross-validation was to estimate the performance of the model, although 19 studies used it for optimising the model-building process in some way (for example in reference S16 it was used to optimise the number of biomarkers to include in the model) and another three used it in the biomarker selection process.

In 24 cases, the cross-validation procedure also included the biomarker selection process.

3.3.2 Permutation testing techniques

18 studies used permutation testing as part of the model-building process. In 13 of these cases, the purpose of the permutation testing was to estimate the significance of the model's performance, whereas in the other five cases the purpose was to estimate the significance of the biomarker-selection process.

The number of permutations used varied considerably between studies, as set out in figure 8 below.

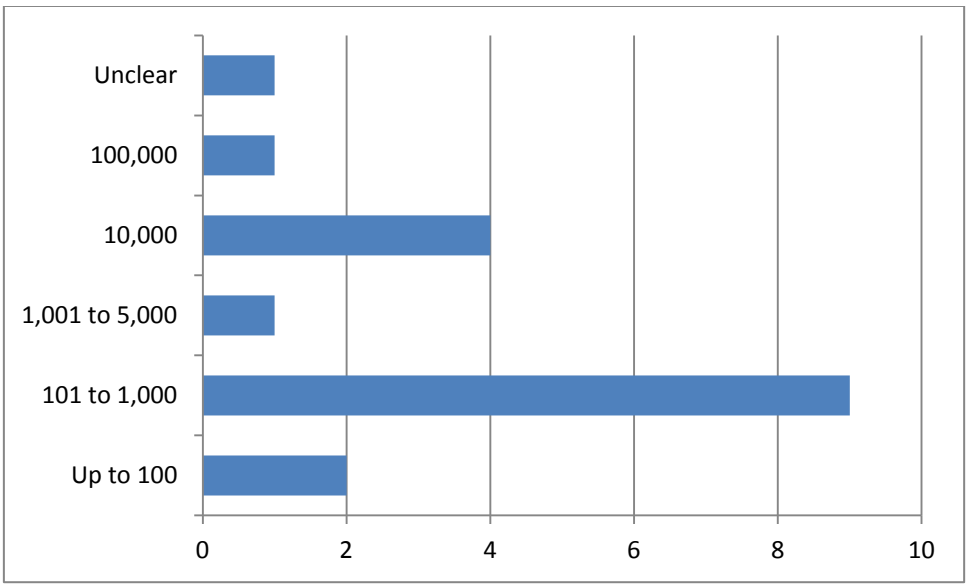


Figure 41 - distribution of numbers of permutations used

3.3.3 Internal validation with separate data

38 studies used an internal validation procedure which was based on separate data from the same source as the data used for model building. This meant that the original data set was split into two, one portion used for model building and the other for model testing. The distribution of the different ways in which this was achieved are set out in figure 9 below.

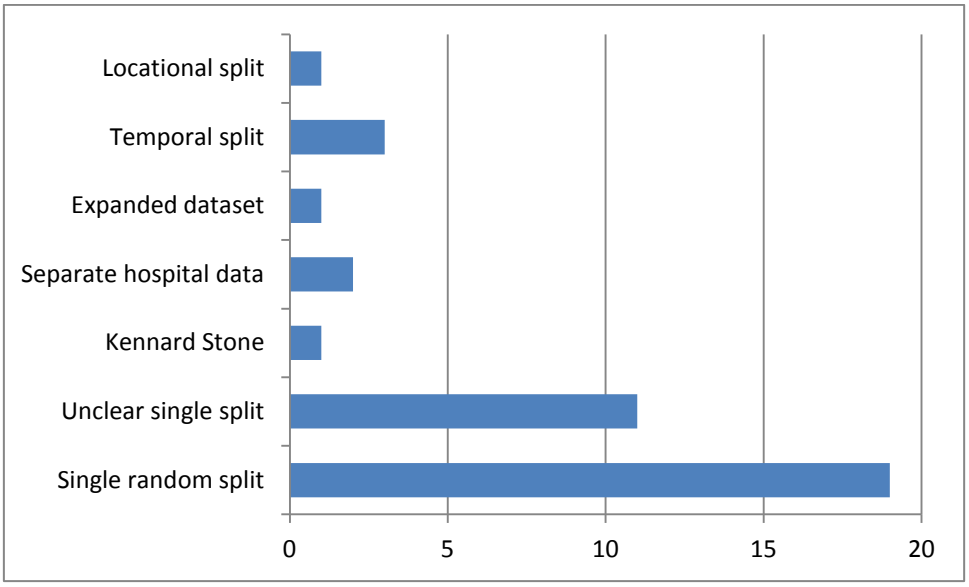


Figure 42 - distribution of methods used for splitting data for internal validation

The most popular method was a single, random split of data, which was used in one half of cases. In 11 cases the approach taken could not be determined.

The number of patients used for validation ranged from 13 to 337, with an average of 76. The number of outcomes amongst these patients ranged from 5 to 96, with an average of 28, in those cases which provided this information. However, in 17 cases the number of outcomes could not be determined.

3.3.4 Comparison of performance results between internal validation methods

Only a small number of studies used more than one method of internal validation. Hence it is difficult to compare how each performed in the same study. An alternative is to look across studies - for example, using classification accuracy as the measure of performance, the average accuracy rate for models which used cross-validation was 87.6%. This compares with an average rate of 78.9% for models which used internal validation with a separate data set.

3.3.5 Summary of key points

Around 75% of studies used some form of internal validation. The most common type of internal validation used was cross-validation, followed by the use of a split data set. For the latter, there was a large range in the number of patients used. Looking across studies, cross-validation tended to result in higher reported model accuracy than internal validation with separate data.

3.4 External validation

62 of the models considered were validated using external datasets. Of these, nine validation procedures were based on prospectively gathered data, whereas 50 were based on retrospective analysis of previously gathered data. The determination for the remaining cases was unclear. Of the nine prospective cases, five appeared to be planned, two opportunistic and for two the determination was unclear.

42 studies which used external validation were based on patient data gathered from different locations and by someone other than the research team who created the model. 44 cases used just one external data set, whilst 10 used two and eight used three or more.

In 17 cases, some patients were removed from the starting external dataset as a result of missing information. The number of patients used ranged from 10 to 1,083, with an average of 194. The number of outcomes amongst these patients ranged from 4 to 112, with an average of 45, in those cases which provided this information. However, in 37 cases the number of outcomes could not be determined.

Summary of key points

Around 60% of studies employed validation using external data. There was a wide range in the numbers of patients used and the numbers of outcomes amongst those patients.

3.5 Comparison of validation and model-building performance measures

The following paragraphs compare the performance of the models when applied to the original data set used for model-building and the validation data sets.

3.5.1 Area under the curve (AUC)

One of the most popular methods for analysing the discriminative ability of a model is the so-called area-under-the-curve (AUC) applied to the receiver operating characteristics (ROC) curve. The ROC curve plots 1-specificity on the x-axis versus sensitivity on the y-axis for all possible model cut-off points. The AUC then gives a measure of how well the model can discriminate cases, with a value of 1 indicating perfect discrimination and a value of 0.5 indicating performance no better than random.

The AUC was reported for 17 of the models created. For six, a corresponding AUC value was reported based on a separate, internal data set, and for seven this was performed on an external data set. Details of the reported AUC measurements are set out in table 6 below.

Ref	Training data		Internal validation data			External validation data		
	Sample size	AUC value	Sample size	AUC value	Change	Sample size	AUC value	Change
S2	61	0.95	-	-	-	24	0.88	-0.07
S29	113	0.87	-	-	-	-	-	-
S32	96	0.87	78	0.80	-0.07	-	-	-
S98	127	0.84	-	-	-	130	0.74	-0.10
S2	50	0.99	-	-	-	-	-	-
S20	122	0.71	172	0.61	-0.10	-	-	-
S35	118	0.75	-	-	-	-	-	-
S49	242	0.76	-	-	-	-	-	-

S59	128	0.77	104	0.74	-0.03	94 & 33	0.61 & 0.68	-0.16 & -0.09
S60	247	0.83	-	-	-	91	0.72	-0.11
S62	153	0.81	-	-	-	-	-	-
S74	110	0.96	-	-	-	87	0.67	-0.29
S7	36	0.87	24	0.82	-0.05	-	-	-
S12	172	0.81	56	0.95	+0.14	-	-	-
S47	30	0.84	69	0.61	-0.23	-	-	-
S87	47	0.89	-	-	-	20	0.77	-0.12
S99	252	0.78	-	-	-	224	0.73	-0.05
Average change relative to training result					**-0.06**			**-0.12**

Table 6 - comparison of AUC values between training and validation datasets

Although the number of cases for which the AUC was reported is small, there is a clear trend showing that the validation AUC is usually lower than that obtained using the training data. For internal validation exercises, the difference between the AUCs reported for validation and training data ranged from -0.23 to +0.14, with an average of -0.06. When external validation was used, the average difference was twice as large at -0.12, with a range of -0.29 to -0.05.

3.5.2 Sensitivity and specificity

In 29 cases the sensitivity and/or specificity of the model applied to the training data was reported. For six of these cases sensitivity and specificity were also reported in a separate internal data set, and for nine both measures were also reported in a separate external data set. Details are set out in table 7 below. Figures in brackets are the change relative to the training data set result.

Ref	Sample sizes			Sensitivity values			Specificity values		
	Training	Internal	External	Training	Internal	External	Training	Internal	External
S2	50	-	-	0.91	-	-	0.75	-	-
S4	20	-	-	0.80	-	-	0.76	-	-
S41	21	-	-	1.00	-	-	1.00	-	-
S49	242	-	-	-	-	-	0.60	-	-
S58	229	-	-	0.73	-	-	0.66	-	-
S59	128	-	-	0.63	-	-	0.81	-	-
S64	160	-	-	0.88	-	-	0.86	-	-
S66	50	34	85	0.77	0.22	0.33	0.86	1.00	0.94

					(-0.55)	(-0.44)		(+0.14)	(+0.08)
S67	103	-	-	0.82	-	-	0.91	-	-
S74	110	-	87	0.89	-	0.64 (-0.25)	0.86	-	0.69 (-0.17)
S79	103	-	-	0.91	-	-	0.94	-	-
S84	56	-	-	0.78	-	-	0.79	-	-
S91	61	-	267	0.64	-	0.62 (-0.02)	0.67	-	0.67 (0)
S95	27	-	-	0.92	-	-	0.93	-	-
S2	61	-	24	0.64	-	1.00 (+0.36)	0.94	-	0.67 (-0.27)
S4	141	-	151	1.00	-	1.00 (0)	1.00	-	0.96 (-0.04)
S30	213	-	42	0.85	-	0.89 (+0.04)	0.81	-	0.54 (-0.27)
S19	44	44	-	0.75	0.91 (+0.16)	-	1.00	0.91 (-0.09)	-
S40	157	80	-	0.94	0.98 (+0.04)	-	0.96	1.00 (+0.04)	-
S41	71	68	-	1.00	1.00 (0)	-	1.00	1.00 (0)	-
S12	172	-	-	0.87	-	-	0.72	-	-
S13	31	-	-	0.79	-	-	0.83	-	-
S18	50	34	-	1.00	0.88 (-0.12)	-	0.59	0.54 (-0.05)	-
S22	57	37	-	0.87	0.73 (-0.14)	-	0.82	0.86 (+0.04)	-
S43	27	-	-	0.77	-	-	0.93	-	-
S45	28	-	14	0.96	-	0.72 (-0.24)	0.80	-	1.00 (+0.20)
S56	25	-	10	0.68	-	0.75 (+0.07)	0.93	-	0.83 (-0.10)
S92	27	-	-	0.75	-	-	0.67	-	-
S99	252	-	224	0.82	-	0.84 (+0.02)	0.63	-	0.52 (-0.11)
Average change relative to training result					**-0.10**	**-0.05**		**+0.01**	**-0.08**

Table 7 - comparison of sensitivity and specificity values between training and validation data sets

The differences between validation and training results are variable (maximum difference of +0.36, lowest of -0.55), in part because some of the sample numbers involved are very small. However, generally speaking there is a trend for the validation results to be less favourable than the training results.

Another way of considering these statistics is to calculate Youden's Index, which equals the sum of sensitivity and specificity minus 1. The index ranges from -1 to 1, with 1 implying perfect predictive ability, 0 implying no predictive ability and -1 implying totally incorrect predictive ability.

For the six cases in table 7 which used internal validation, Youden's Index value reduced on average by 0.09 between the training and validation data sets (and in fact it actually increased in two cases). For the nine external validation cases, the average reduction was greater at 0.13, and only one case resulted in an increase.

3.5.3 Calibration

Only one paper (ref S58) included any information on model calibration, and even then it was only shown for the combined training and validation data sets with no split between them.

3.5.4 Decision analysis

None of the 100 papers included any information on cost / benefit analysis of applying the model in practice.

3.5.5 Summary of key points

Relevant performance measures were presented in only a small proportion of studies. When they were presented, results using validation data tended to be worse than those based on the data used to create the model.

4. Discussion

The analysis of the 100 papers has highlighted a number of issues concerning the way in which validation studies are undertaken.

Firstly, only 5% of the papers were purely concerned with the validation of a previously constructed model, with the vast majority covering both the construction and subsequent validation of a model by the same research team. This phenomenon has previously been observed[6]. Whilst it is understandable that individuals may prefer to concentrate on creating their own models rather than validating someone else's work, it is worrying that little effort seems to be being put into genuine, external validation of models created by other researchers.

It is encouraging that over 50% of studies used external validation. However, it is of concern that internal validation is more common than external validation given that the former is considered to be prone to overstating a model's predictive ability[5]. Again this trend for more frequent internal validation has been observed in previous research[6], and is understandable when suitable data cannot be found in adequate quantity. The tendency for internal validation to produce more favourable measures of performance than external validation was again observed in the small number of cases for which a direct comparison could be made.

There are a couple of particular concerns in connection with internal validation practices. Firstly, cross-validation was the most popular method of internal validation. In order to minimise the risk of bias in performance measurement, it is important that cross-validation covers the entire model-building process, including the selection of biomarkers[7]. However, this part of the process was covered in less than half of the cross-validation exercises encountered in the review.

Secondly, for those cases where internal validation involved the split of the data set into training and test sets, there was a wide range of sample sizes used. The average number of patients was 76, and the average number of outcomes was only 28. The typical ratio of outcomes to biomarkers in the model was much lower than the figure of 10 which is considered to be a reasonable minimum for acceptable statistical power[6]. Only three studies met this criterion. These small sample sizes bring into question the ability of the validation exercises to adequately assess model performance. Previous research[8] has indicated that, in certain circumstances, the rule of thumb may be relaxed below 10 outcomes per biomarker, but even then the analysis in this study suggests that in many cases the number of outcomes is still inadequate.

Another concern is that in 45% of internal validation cases the number of outcomes amongst the patient population used could not be determined. This too is a phenomenon which has been observed before[9], and means that the standard of reporting often does not meet the REMARK guidelines[10], which were established to encourage transparent reporting in prognostic studies involving tumour biomarkers.

The issue of inadequate sample sizes and unclear reporting of numbers of outcomes was also a concern for those studies which used external validation. Previous research has suggested that for effective external validation, the validation data should include at least 100 outcomes and another 100 non-outcomes[11]. The analysis in this project has shown that the numbers used, in particular for outcomes (where this could be determined), were generally a lot lower than this benchmark, with only two studies meeting it.

This study has also identified that in over 25% of external validation exercises, some patients were removed from the initial validation data set because of incomplete data. The issue here is that it is unlikely that missing data is a random event, but rather that it is linked in some way to characteristics of the disease in question, and therefore possibly to the chance of an outcome occurring. This means that excluding patients because of missing data may lead to bias in the performance measurement[12].

Linked to the above issue is the fact that over 80% of external validation exercises used retrospectively gathered data. Retrospective studies have the potential to introduce bias into the results because of the availability of tumour material[13]. Again, the overuse of retrospective studies is a phenomenon which has previously been observed in the context of reporting of methods used in developing prognostic models[14]. Another observation is that the same retrospective data has often been used in the creation and/or validation of several different models - for example, the same data set was used in five separate studies involving lung cancer. Whilst one can understand this from the practical point of view of obtaining suitable data, the use of the same data set for different studies reduces the independence of these studies.

Another observation relating to data is that in 7% of studies the validation data was actually used in some part of the model-building process. This is an unfortunate practice as it violates the independence of the validation data from the model, and can therefore lead to "information leak" which can compromise the accuracy of the validation results[1,7].

Turning to the reporting of performance measurements, the use of standard statistical techniques (such as AUC) for reporting discrimination was uncommon. Much greater emphasis was placed on reporting Kaplan-Meier estimates of survival, with associated p-values, and also on hazard ratios from analyses using Cox proportional hazards. This in turn followed on from the very common approach of dichotomising patients into two risk-groups, rather than treating model output as a continuous variable (which in most cases it was). This is generally viewed as a bad idea[15], as it introduces an artificial, arbitrary cut-off point which has no biological justification and can lead to biased performance measurements.

Calibration was only reported in one study. Perhaps unsurprisingly, none of the studies included any detailed analysis of how the model could impact actual clinical decision making.

Given that only a very small proportion of biomarker models are eventually adopted into clinical practice[2], a review of the way in which such models are created and validated would seem necessary.

Probably the most important issue is that of scale - i.e. using very large sample sizes, for both model creation and subsequent external validation. This will require more collaboration and sharing of data between research teams (ideally on a worldwide basis), rather than the present system of individual research teams working on their own models with limited access to data (a lot of which is retrospectively gathered). Such collaboration may be difficult to coordinate and expensive, but then there currently seems to be a lot of time and effort wasted on producing models which will never get anywhere near being used in practice.

As well as greater collaboration, more emphasis needs to be placed on proper validation. This could be achieved, for example, through the peer-review process when papers are submitted for publication.

In conclusion, even this relatively small study has identified several areas of concern with the way in which validation studies have been conducted in recent years. Many of these issues, in particular those relating to inadequate sample sizes and lack of transparency in reporting data, were also present in the model-building process as well as validation.

As well as the small size, another key limitation of this study is that all the work was performed by one individual. Ideally at least two people would have reviewed all the papers and then compared their analyses, thereby providing a check on the results. In many of the papers reviewed there was a lack of clarity in reporting, which meant that judgements had to be made as to how to answer certain questions. It would have been preferable for these judgements to be performed by more than one person, to reduce the possibility of personal bias entering the analysis.

Appendix - 100 papers included in survey

S1 Yu K-D, Huang A-J, Fan L, Li W-F and Shao Z-M (2012) Genetic variants in oxidative stress-related genes predict chemoresistance in primary breast cancer: a prospective observational study and validation. Cancer Research 72(2), 408-419

S2 Martens-Uzunova ES, Jalava SE, Dits NF, van Leenders GJLH, Moller S et al (2012) Diagnostic and prognostic signatures from the small non-coding RNA transcriptome in prostate cancer. Oncogene 31, 978-991

S3 Patel JP, Gonen M, Figueroa ME, Fernandez M, Sun Z et al (2012) Prognostic relevance of integrated genetic profiling in acute myeloid leukemia. The New England Journal of Medicine 366(12), 1079-1089

S4 Bertini I, Cacciatore S, Jensen BV, Schou JV, Johansen JS et al (2012) Metabolomic NMR fingerprinting to identify and predict survival of patients with metastatic colorectal cancer. Cancer Research 72(1), 356-364

S5 Stretch C, Eastman T, Mandal R, Eisner R, Wishart DS et al (2012) Prediction of skeletal muscle and fat mass in patients with advanced cancer using a metabolomic approach. The Journal of Nutrition 142, 14-21

S6 Xie Y, Xiao G, Coombes KR, Behrens C, Solis LM et al (2011) Robust gene expression signature from formalin-fixed paraffin-embedded samples predicts prognosis of non-small-cell lung cancer patients. Clinical Cancer Research 17(17), 5705-5714

S7 Kerr DA and Wittliff JL (2011) A five-gene model predicts clinical outcome in ER+/PR+ early-stage breast cancers treated with adjuvant tamoxifen. Hormones and Cancer 2, 261-271

S8 Hu J, Wang Z, Fan J, Dai Z, He Y-F et al (2011) Genetic variations in plasma circulating DNA of HBV-related hepatocellular carcinoma patients predict recurrence after liver transplantation. PLoS One 6(10), e26003

S9 Chen D-T, Hsu Y-L, Fulp WJ, Coppola D, Haura EB et al (2011) Prognostic and predictive value of a malignancy-risk gene signature in early-stage non-small cell lung cancer. Journal of National Cancer Institute 103, 1859-1870

S10 Reis PP, Waldron L, Perez-Ordonez B, Pintilie M, Galloni NN et all (2011) A gene signature in histologically normal surgical margins is predictive of oral carcinoma recurrence. BMC Cancer 11:437

S11 Hao K, Lamb J, Zhang C, Xie T, Wang K et al (2011) Clinicopathologic and gene expression parameters predict liver cancer prognosis. BMC Cancer 11:481

S12 Wu P, Walker BA, Brewer D, Gregory WM, Ashcroft J et al (2011) A gene expression-based predictor for myeloma patients at high risk of developing bone disease on bisphosphonate treatment. Clinical Cancer Research 17(19), 6347-6355

S13 Maher SG, McDowell DT, Collins BC, Muldoon C, Gallagher WM et al (2011) Serum proteomic profiling reveals that pretreatment complement protein levels are predictive of esophageal cancer patient response to neoadjuvant chemoradiation. Annals of Surgery 254, 809-817

S14 Miller LD, Coffman LG, Chou JW, Black MA, Bergh J et al (2011) An iron regulatory gene signature predicts outcome in breast cancer. Cancer Research 71(21), 6728-6737

S15 Cuzick J, Dowsett M, Pineda S, Wale C, Salter J et al (2011) Prognostic value of a combined estrogen receptor, progetserone receptor, Ki-67, and human epidermal growth factor receptor 2 immunohistochemical score and comparison with the genomic heath recurrence score in early breast cancer. Journal of Clinical Oncology 29, 4273-4278

S16 Toustrup K, Sorensen BS, Nordsmark M, Busk M, Wiuf C et al (2011) Development of a hypoxia gene expression classifier with predictive impact for hypoxic modification of radiotherapy in head and neck cancer. Cancer Research 71(17), 5923-5931

S17 Rinaldi A, Mian M, Kwee I, Rossi D, Deambrogi C et al (2011) Genome-wide DNA profiling better defines the prognosis of chronic lymphocytic leukaemia. British Journal of Haematology 154, 590-599

S18 Naoi Y, Kishi K, Tanei T, Tsunashima R, Tominaga N et al (2011) Prediction of pathologic complete response to sequential paclitaxel and 5-fluorouracil/epirubicin/cyclophosphamide therapy using a 70-gene classifier for breast cancers. Cancer 117, 3682-3690

S19 Huang L, Zheng M, Zhou Q-M, Zhang M-Y, Jia W-H et al (2011) Identification of a gene-expression signature for predicting lymph node metastasis in patients with early stage cervical carcinoma. Cancer 117, 3363-3373

S20 Gao Q, Wang X-Y, Qiu S-J, Zhou J, Shi Y-H et al (2011) Tumor stroma reaction-related gene signature predicts clinical outcome in human hepatocellular carcinom. Cancer Science 102, 1522-1531

S21 Sabatier R, Finetti P, Bonensea J, Jacquemier J, Adelaide J et al (2011) A seven-gene prognostic model for platinum-treated ovarian carcinomas. British Journal of Cancer 105, 304-311

S22 Cassado E, Garcia VM, Sanchez JV, Blanco M, Maurel J et al (2011) A combined strategy of SAGE and quantitative PCR provides a 13-gene signature that predicts preoperative chemoradiotherapy response and outcome in renal cancer. Clinical Cancer Research 17(12), 4145-4154

S23 Kim W-J, Kim S-J, Jeong P, Yun S-K, Cho I-C (2011) A four-gene signature predicts disease progression in muscle invasive bladder cancer. Molecular Medicine 17(5-6), 478-485

S24 Hatzis C, Pusztai L, Valero V, Booser DJ, Esserman L et al (2011) A genomic predictor of response and survival following taxane-anthracycline chemotherapy for invasive breast cancer. JAMA 305(18), 1873-1881

S25 Bertucci F, Borie N, Roche H, Bachelot T, le Doussal J-M et al (2011) Gene expression profile predicts outcome after anthracycline-based adjuvant chemotherapy in early breast cancer. Breast Cancer Research and Treatment 127, 363-373

S26 Mitra R, Lee J, Jo J, Milani M, McClintick JN et al (2011) Prediction of post-operative recurrence-free survival in non-small cell lung cancer by using an internationally validated gene expression model. Clinical Cancer Research 17(9), 2934-2946

S27 Navab R, Strumpf D, Bandarchi B, Zhu C-Q, Pintilie M et al (2011) Prognostic gene-expression signature of carcinoma-associated fibroblasts in non-small cell lung cancer. PNAS 108(17), 7160-7165

S28 Paulson KG, Iyer JG, Tegeder AR, Thibodeau R, Schelter J et al (2011) Transcriptome-wide studies of Merkel cell carcinoma and validation of intratumural CD8+ lymphocyte invasion as an independent predictor of survival. Journal of Clinical Oncology 29, 1539-1546

S29 Mendez E, Lohavanichbutr P, Fan W, Houck JR, Rue TC et al (2011) Can a metastatic gene expression profile outperform tumor size as a predictor of occult lymph node metastasis in oral cancer patients? Clinical Cancer Research 17(8), 2466-2473

S30 Coutant C, Rouzier R, Qi Y, Lehmann-Che J, Bianchini G et al (2011) Distinct p53 gene signatures are needed to predict prognosis and response to chemotherapy in ER-positive and ER-negative breast cancers. Clinical Cancer Research 17(8), 2591-2601

S31 Zhang Y-Z, Zhang L-H, Gao Y, Li C-H, Jia S-Q et al (2011) Discovery and validation of prognostic markers in gastric cancer by genome-wide expression profiling. World Journal of Gastroenterology 17(13), 1710-1717

S32 Tamayo P, Cho Y-J, Tsherniak A, Greulich H, Ambrogio L et al (2011) Predicting relapse in patients with medulloblastoma by integrating evidence from clinical and

genomic features. Journal of Clinical Oncology 29, 1415-1423

S33 Cho JY, Lim JY, Cheong JH, Park Y-Y, Yoon S-L et al (2011) Gene expression signature-based prognostic risk score in gastric cancer. Clinical Cancer Research 17(7), 1850-1857

S34 Gobble RM, Qin L-X, Brill ER, Angeles CV, Ugras S et al (2011) Expression profiling of liposarcoma yields a multigene predictor of patient outcome and identifies genes that contribute to liposarcomagenesis. Cancer Research 71(7), 2697-2705

S35 Zhao X, Rodland EA, Sorlie T, Naume B, Langerod A et al (2011) Combining gene signatures improves prediction of breast cancer survival. PLoS One 6(3): e17845

S36 Yi JM, Dhir M, Van Neste L, Downing SR, Jeschke J et al (2011) Genomic and epigenomic integration identifies a prognostic signature in colon cancer. Clinical Cancer Research 17(6), 1535-1545

S37 Sabatier R, Finetti P, Cervera N, Lambaudie E, Esterni B et al (2011) A gene expression signature identifies two prognostic subgroups of basal breast cancer. Breast Cancer Research and Treatment 126, 407-420

S38 Rossi D, Rasi S, Di Rocco A, Fabbri A, Forconi F et al (2011) The host genetic background of DNA repair mechanisms is an independent predictor of survival in diffuse large B-cell lymphoma. Blood 117(8), 2405-2413

S39 Saintigny P, Zhang L, Fan Y-H, El-Naggar AK, Papadimitrakopoulou VA et al (2011) Gene expression profiling predicts the development of oral cancer. Cancer Prevention Research 4(2), 218-229

S40 Balgobind BV, Van den Heuvel-Elbrink MM, De Menezes RX, Reinhardt D, Hollink IHIM et al (2011) Evaluation of gene expression signatures predictive of cytogenetic and molecular subtypes of pedriatic acute myeloid leukemia. Haematologica 96(2), 221-230

S41 Gotoh M, Arai E, Wakai-Ushijima S, Hiraoka N, Kosuge T et al (2011) Diagnosis and prognostication of ductal adenocarcinomas of the pancreas based on genome-wide DNA methylation profiling by bacterial artificial chromosome array-based methylated CpG island amplification. Journal of Biomedicine and Biotechnology 2011, 780836

S42 Salazar R, Roepman P, Capella G, Moreno V, Simon I et al (2010) Gene expression signature to improve prognosis prediction of stage II and III colorectal cancer. Journal of Clinical Oncology 29, 17-24

S43 Li Y, Dang TA, Shen J, Hicks J, Chintagumpala M et al (2011) Plasma proteome predicts chemotherapy response in osteosarcoma patients. Oncology Reports 25, 303-314

S44 Onozato W, Yamashita K, Yamahshita K, Kuba T, Katoh H et al (2011) Genetic alterations of K-ras may reflect prognosis in stage III colon cancer patients below 60 years of age. Journal of Surgical Oncology 103, 25-33

S45 Barros Filho MC, Katayama MLH, Brentani H, Abreu APS, Barbosa EM et al (2010) Gene trio signatures as molecular markers to predict response to doxorubicin cyclophosphamide neoadjuvant chemotherapy in breast cancer patients. Brazilian Journal of Medical and Biological Research 43, 1225-1231

S46 Tong D, Heinze G, Pils D, Wolf A, Singer CF et al (2010) Gene expression of PMP22 is an independent prognostic factor for disease-free and overall survival in breast cancer patients. BMC Cancer 10:682

S47 Paik H, Lee E and Lee D (2010) Relationships between genetic polymorphisms and transcriptional profiles for outcome prediction in anticancer agent treatment. BMB Reports 43(12), 836-841

S48 Daghistani M, Marin D, Khorashad JS, Wang L, May PC et al (2010) EVI-1 oncogene expression predicts survival in chronic-phase CML patients resistant to imatinib treated with second-generation tyrosine kinase inhibitors. Blood 116(26), 6014-6017

S49 Roessler S, Jia H-L, Budhu A, Forgues M, Ye Q-H et al (2010) A unique metastasis gene signature enables prediction of tumor relapse in early-stage hepatocellular carcinoma patients. Cancer Research 70(24), 10202-10212

S50 Kim SM, Park Y-Y, Park ES, Cho JY, Izzo JG et al (2010) Prognostic biomarkers for esophageal adenocarcinoma identified by analysis of tumor transcriptome. PLoS One 5(11): e15074

S51 Nguyen GH, Schetter AJ, Chou DB, Bowman ED, Zhao R et al (2010) Inflammatory and MicroRNA gene expression as prognostic classifier of Barrett's-Associated esophageal adenocarcinoma. Clinical Cancer Research 16(23), 5824-5834

S52 Peters CJ, Rees JRE, Hardwick RH, Hardwick JS, Vowler SL et al (2010) A 4-gene signature predicts survival of patients with resected adenocarcinoma of the esophagus, junction, and gastric cardia. Gastroenterology 139, 1995-2004

S53 Kawarazaki S, Taniguchi K, Shirahata M, Kukita Y, Kanemoto M et al (2010) Conversion of a molecular classifier obtained by gene expression profiling into a classifier based on real-time PCR: a prognosis predictor for gliomas. BMC Medical Genomics 3:52

S54 Zhu C-Q, Strumpf D, Li C-Y, Li Q, Liu N et al (2010) Prognostic gene expression signature for squamous cell carcinoma of lung. Clinical Cancer Research 16(20),

5038-5047

S55 Bianco-Miotto T, Chiam K, Buchanan G, Jindal S, Day TK et al (2010) Global levels of specific histone modifications and an epigenetic gene signature predict prostate cancer progression and development. Cancer Epidemiology, Biomarkers and Prevention 19(10), 2611-2622

S56 Motoori M, Takemasa I, Yamasaki M, Komori T, Takeno A et al (2010) Prediction of the response to chemotherapy in advanced esophageal cancer by gene expression profiling of biopsy samples. International Journal of Oncology 37, 1113-1120

S57 Lin Y, Lin S, Watson M, Trinkaus KM, Kuo S et al (2010) A gene expression signature that predicts the therapeutic response of the basal-like breast cancer to neoadjuvant chemotherapy. Breast Cancer Research and Treatment 123, 691-699

S58 Wan Y-W, Sabbagh E, Raese R, Qian Y, Luo D et al (2010) Hybrid models identified a 12-gene signature for lung cancer prognosis and chemoresponse prediction. PLoS One 5(8):e12222

S59 Kim J, Hong SJ, Park JY, Park JH, Yu Y-S et al (2010) Epithelial-mesenchymal transition gene signature to predict clinical outcome of hepatocellular carcinoma. Cancer Science 101(6), 1521-1528

S60 Van Malenstein H, Gevaert O, Libbrecht L, Daemen A, Allemeersch J et al (2010) A seven-gene set associated with chronic hypoxia of prognostic importance in hepatocellular carcinoma. Clinical Cancer Research 16(16), 4278-4288

S61 Stratford JK, Bentrem DJ, Anderson JM, Fan C, Volmar KA et al (2010) A six-gene signature predicts survival of patients with localized pancreatic ductal adenocarcinoma. PLoS Medicine 7(7):e1000307

S62 Sanchez-Navarro I, Gamez-Pozo A, Pinto A, Hardisson D, Madero R et al (2010) An 8-gene qRT-PCR-based gene expression score that has prognostic value in early breast cancer. BMC Cancer 10:336

S63 Lee H-J, Nam KT, Park HS, Kim MA, Lafleur BJ et al (2010) Gene expression profiling of metaplastic lineages identifies CDH17 as a prognostic marker in early stage gastric cancer. Gastroenterology 139, 213-225

S64 Chibon F, Lagarde P, Salas S, Perot G, Brouste V et al (2010) Validated prediction of clinical outcome in sarcomas and multiple types of cancer on the basis of a gene expression signature related to genome complexity. Nature Medicine 16(7), 781-788

S65 Watanabe T, Kobunai T, Yamamoto Y, Kanazawa T, Konishi T et al (2010) Prediction of liver metastasis after colorectal cancer using reverse transcription-polymerase chain reaction analysis of 10 genes. European Journal of Cancer 46, 2119-2126

S66 Cleaver AL, Beesley AH, Firth MJ, Sturges NC, O'Leary RA et al (2010) Gene-based outcome prediction in multiple cohorts of pediatric T-cell acute lymphoblastic leukemia: a Children's Oncology Group study. Molecular Cancer 9:105

S67 Lee J-S, Leem S-H, Lee S-Y, Kim S-C, Park E-S et al (2010) Expression signature of E2F1 and its associated genes predict superficial to invasive progression of bladder tumors. Journal of Clinical Oncology 28, 2660-2667

S68 Hou J, Aerts J, den Hamer B, van IJcken W, den Bakker M et al (2010) Gene expression-based classification of non-small cell lung carcinomas and survival prediction. PLoS One 5(4):e10312

S69 Chng WJ, Gertz MA, Chung T-H, Van Wier S, Keats JJ et al (2010) Correlation between array-comparative genomic hybridization-defined genomic gains and losses and survival: identification of 1p31-32 deletion as a prognostic factor in myeloma. Leukemia 24, 833-842

S70 Staaf J, Ringner M, Vallon-Christersson J, Jonsson G, Bendahl P-O et al (2010) Identification of subtypes in human epidermal growth factor receptor 2-positive breast cancer reveals a gene signature prognostic of outcome. Journal of Clinical Oncology 28, 1813-1820

S71 Mollerstrom E, Delle U, Danielsson A, Parris T, Olsson B et al (2010) High-resolution genomic profiling to predict 10-year overall survival in node-negative breast cancer. Cancer Genetics and Cytogenetics 198, 79-89

S72 Kim Y-J, Ha Y-S, Kim S-K, Yoon HY, Lym MS et al (2010) Gene signatures for the prediction of response to Bacillus Calmette-Guerin immunotherapy in primary pT1 bladder cancers. Clinical Cancer Research 16(7), 2131-2137

S73 Hu Z, Chen X, Zhao Y, Tian T, Jin G et al (2010) Serum microRNA signatures identified in a genome-wide serum microRNA expression profiling predict survival of non-small-cell lung cancer. Journal of Clinical Oncology 28, 1721-1726

S74 Yoshihara K, Tajima A, Yahata T, Kodama S, Fujiwara H et al (2010) Gene expression profile for predicting survival in advanced-stage serous ovarian cancer across two independent datasets. PLoS One 5(3):e9615

S75 Takeno A, Takemasa I, Seno S, Yamasaki M, Motoori M et al (2010) Gene expression profile prospectively predicts peritoneal relapse after curative surgery of gastric cancer. Annals of Surgical Oncology 17, 1033-1042

S76 Maeshima H, Ohno K, Nakano S and Yamada T (2010) Validation of an in vitro screening test for predicting the tumor promoting potential of chemicals based on gene expression. Toxicology in Vitro 24, 995-1001

S77 Smith JJ, Deane NG, Wu F, Merchant NP, Zhang B et al (2009) Experimentally derived metastasis gene expression profile predicts recurrence and death in patients with colon cancer. Gastroenterology 138, 958-968

S78 Davicioni E, Anderson JR, Buckley JD, Meyer WH and Triche TJ (2010) Gene expression profiling for survival prediction in pediatric rhabdomyosarcomas: a report from the Children's Oncology Group. Journal of Clinical Oncology 28, 1240-1246

S79 Kim S-Y, Kim E-J, Leem S-H, Ha Y-S, Kim Y-J and Kim W-J (2010) Identification of S100A8-correlated genes for prediction of disease progression in non-muscle invasive bladder cancer. BMC Cancer 10:21

S80 Kang H, Chen I-M, Wilson CS, Bedrick EJ, Harvey RC et al (2010) Gene expression classifiers for relapse-free survival and minimal residual disease improve risk classification and outcome prediction in pediatric B-precursor acute lymphoblastic leukemia. Blood 115, 1394-1405

S81 Kim W-J, Kim E-J, Kim S-K, Kim Y-J, Ha Y-S et al (2010) Predictive value of progression-related gene classifier in primary non-muscle invasive bladder cancer. Molecular Cancer 9:3

S82 Giskeodegard GF, Grinde MT, Sitter B, Axelson DE, Lundgren S et al (2010) Multivariate modeling and prediction of breast cancer prognostic factors using MR metabolomics. Journal of Proteome Research 9, 972-979

S83 McWeeney SK, Pemberton LC, Loriaux MM, Vartanian K, Willis SG et al (2010) A gene expression signature of CD34+ cells to predict majr cytogenetic response in chronic-phase chronic myeloid leukemia patients treated with imatinib. Blood 115, 315-325

S84 Baty F, Facompre M, Kaiser S, Schumacher M, Pless M et al (2010) Gene profiling of clinical routine biopsies and prediction of survival in non-small cell lung cancer. American Journal of Respiratory and Critical Care Medicine 181, 181-188

S85 Paris PL, Weinberg V, Albo G, Roy R, Burke C et al (2010) A group of genome-based biomarkers that add to a Kattan Nomogram for predicting progression in men with high-risk prostate cancer. Clinical Cancer Research 16(1), 195-202

S86 Banelli B, Bonassi S, Casciano I, Mazzocco , Di Vinci A et al (2009) Outcome prediction and risk assessment by quantitative pyrosequencing methylation analysis of the SFN gene in advanced stage, high-risk, neuroblastic tumor patients. International Journal of Cancer 126, 656-668

S87 Lee S-C, Xu X, Chng W-J, Watson M, Lim Y-W et al (2009) Post-treatment tumor gene expression signatures are more predictive of treatment outcomes than baseline

signatures in breast cancer. Pharmacogenetics and Genomics 19, 833-842

S88 Gravendeel LAM, Kouwenhoven MCM, Gevaert O, de Rooi JJ, Stubbs AP et al (2009) Intrinsic gene expression profiles of gliomas are a better predictor of survival than histology. Cancer Research 69(23), 9065-9072

S89 Hsu Y-C, Yuan S, Chen H-Y, Yu S-L, Liu C-H et al (2009) A four-gene signature from NCI-60 cell line for survival prediction in non-small cell lung cancer. Clinical Cancer Research 15(23), 7309-7315

S90 Gould Rothberg BE, Berger AJ, Molinaro AM, Subtil A, Krauthammer MO et al (2009) Melanoma prognostic model using tissue microarrays and genetic algorithms. Journal of Clinical Oncology 27, 5772-5780

S91 Friedman DR, Weinberg JB, Barry WT, Goodman BK, Volkheimer AD et al (2009) A genomic approach to improve prognosis and predict therapeutic response in chronic lymphocytic leukemia. Clinical Cancer Research 15(22), 6947-6955

S92 Maher SG, Gillham CM, Duggan SP, Smyth PC, Miller N et al (2009) Gene expression analysis of diagnostic biopsies predicts pathological response to neoadjuvant chemoradiotherapy of esophageal cancer. Annals of Surgery 250, 729-737

S93 Korkola JE, Houldsworth J, Feldman DR, Olshen AB, Qin L-X et al (2009) Identification and validation of a gene expression signature that predicts outcome in adult men with germ cell tumors. Journal of Clinical Oncology 27, 5240-5247

S94 Specht K, Harbeck N, Smida J, Annecke K, Reich U et al (2009) Expression profiling identifies genes that predict recurrence of breast cancer after adjuvant CMF-based chemotherapy. Breast Cancer Research and Treatment 118, 45-56

S95 Kim M and Chung HC (2009) Standardized genetic alteration score and predicted score for predicting recurrence status of gastric cancer. Journal of Cancer Research and Clinical Oncology 135, 1501-1512

S96 Bueno-de-Mesquita JM, Linn SC, Keijzer R, Wesseling J, Nuyten DSA et al (2009) Validation of 70-gene prognosis signature in node-negative breast cancer. Breast Cancer Research and Treatment 117, 483-495

S97 Mitra AP, Pagliarulo V, Yang D, Waldman FM, Datar RH et al (2009) Generation of a concise gene panel for outcome prediction in urinary bladder cancer. Journal of Clinical Oncology 27, 3929-3937

S98 Schiffer E, Vlahou A, Petrolekas A, Stravodimos K, Tauber R et al (2009) Prediction of muscle-invasive bladder cancer using urinary proteomics. Clinical Cancer Research 15(15), 4935-4943

S99 Jezequel P, Campone M, Roche H, Gouraud W, Charbonnel C et al (2009) A 38-gene expression signature to predict metastasis risk in node-positive breast cancer after systemic adjuvant chemotherapy: a genomic substudy of PACS01 clinical trial. Breast Cancer Research and Treatment 116, 509-520

S100 Weidhaas JB, Li S-X, Winter K, Ryu J, Jhingran A et al (2009) Changes in gene expression predicting local control in cervical cancer: results from Radiation Therapy Oncology Group 0128. Clinical Cancer Research 15(12), 4199-4206

List of references

1. Taylor JMG, Ankerst DP and Andridge RR (2008) Validation of biomarker-based risk prediction models. Clinical Cancer Research 14(19), 5977-5983

2. Ioannidis JPA and Khoury MJ (2011) Improving validation practices in "Omics" research. Science 334, 1230-1232

3. Altman DG, Vergouwe Y, Royston P and Moons KGM (2009) Prognosis and prognostic research: validating a prognostic model. BMJ 338, 1432-1435

4. Moons KGM, Royston P, Vergouwe Y, Grobbee DE and Altman DG (2009) Prognosis and prognostic research; what, why and how? BMJ 338, 1317-1320

5. Bleeker SE, Moll HA, Steyerberg EW, Donders ART, Derksen-Lubsen G et al (2003) External validation is necessary in prediction research: a clinical example. Journal of Clinical Epidemiology 56, 826-832

6. Bouwmeester W, Zuithoff NPA, Mallett S, Geerlings MI, Vergouwe Y et al (2012) Reporting and methods in clinical prediction research: a systematic review. PLoS Medicine 9(5), 1-12

7. Castadi PJ, Dahabreh IJ and Ioannidis JPA (2011) An empirical assessment of validation practices for molecular classifiers. Briefings in bioinformatics 12(3), 189-202

8. Vittinghoff E and McCulloch CE (2007) Relaxing the rule of ten events per variable in logistic and cox regression. American Journal of Epidemiology 165(6), 710-718

9. Mallett S, Timmer A, Sauerbrei W, Altman DG (2010) Reporting of prognostic studies of tumour markers: a review of publised articles in relation to REMARK guidelines. British Journal of Cancer 102, 173-180

10. McShane LM, Altman DG, Sauerbrei W, Taube SE, Gion M et al (2005) REporting recommendations for tumour MARKer prognostic studies (REMARK). British Journal of Cancer 93, 387-391

11. Vergouwe Y, Steyerberg EW, Eijkemans MJC and Habbema JDF (2005) Substantial effective sample sizes were required for external validation studies of predictive logistic regression models. Journal of Clinical Epidemiology 58, 475-483

12. Royston P, Moons KGM, Altman DG and Vergouwe Y (2009) Prognosis and prognostic research: Developing a prognostic model. BMJ 338, 1373-1377

13. Hoppin JA, Tolbert PE, Taylor JA, Schroeder, JC and Holly EA (2002) Potential for selection bias with tumour tissue retrieval in molecular epidemiology studies. Annals of Epidemiology 12, 1-6

14. Mallett S, Royston P, Dutton S, Waters R and Altman DG (2010) Reporting methods in studies developing prognostic models in cancer: a review. BMC Medicine 8:20, 1-11

15. Royston P, Altman DG and Sauerbrei W (2006) Dichotomizing continuous predictors in multiple regression: a bad idea. Statistics in Medicine 25, 127-141